Automated News Categorization Using Natural Language Processing

Dharanidhar Reddy Banala (016671226), Nitish Reddy Gaddam (016063242),
Naga Pavan Kumar Kodeboina (016646396)
Department of Applied Data Science, San Jose State University (SJSU)
Data 240: Data Mining and Analytics

MOTIVATION AND BACKGROUND

In today's digital era, the traditional methods of manually sorting and classifying news content have become impractical and time-consuming. The increasing reliance on the internet as a primary news source has amplified the requirement for automatic categorization. This automated system is designed to serve multiple stakeholders: news outlets striving for better content organization, businesses aiming to monitor media coverage, and individuals seeking tailored news feeds. By leveraging advanced Natural Language Processing (NLP) techniques, the project endeavors to develop an intelligent system that can accurately and swiftly categorize news articles into their respective topics, thereby enhancing accessibility and comprehensibility. Over the years, NLP has experienced significant advancements, notably with the advent of machine learning and deep learning techniques. These developments have paved the way for sophisticated applications such as speech recognition, sentiment analysis, language translation, and, pertinent to this project, text classification. The project harnesses these capabilities to focus specifically on news articles, which present unique challenges due to their diverse topics, styles, and structures. The exponential growth of online news content has made manual classification unfeasible, necessitating an automated approach. This project, therefore, addresses the urgent need for an efficient and accurate automated news categorization system.

LITERATURE REVIEW

A. Rizaldy and H. A. Santoso (2017) contributed to the field of news categorization by employing the SVM method for text classification, a machine learning approach that has been successfully applied in previous studies. The integration of SVM and Information Gain, particularly in the context of the Indonesian language, demonstrated notable improvements, yielding a satisfactory accuracy of 98.06%. Usmani and J. A. Shamsi (2020) proposed a scheme tailored for the specific context of filtering and categorizing news headlines related to the Pakistan Stock Exchange (PSX) with minimal manual effort. The authors employ a supervised classification technique to ensure effective segregation of news categories and validate the results through an Artificial Neural Network (ANN) based multiclass classification. U. Suleymanov et.al (2018) introduced a text classification system designed to automate the news classification process, specifically tailored for Azerbaijani news articles. The authors employed Naive Bayes, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) and results indicate that the developed system is capable of handling complex classification tasks effectively, with ANN outperforming SVM for their specific datasets. Liu et.al. (2022) proposed a cutting-edge framework that integrates bidirectional long short-term memory (Bi-LSTM) with the transformer model to enhance classification accuracy. In this innovative framework, the self-attention mechanism of the transformer is replaced with Bi-LSTM, enabling the capture of semantic information from sentences. An attention mechanism is additionally employed to focus on crucial words, adjusting their weights to mitigate long-distance information loss. The incorporation of a pooling network reduces network complexity and highlights main features by halving the dimension of the hidden state.

The success of our news categorization project employing logistic regression, support vector classifier, and random forest models can be attributed to the careful consideration of these classifiers' strengths and their suitability for the specific task at hand. While the literature reviews highlighted the efficacy of methods like SVM and deep learning in various contexts, our choice of models aligns with the characteristics of our dataset and the nature of news categorization. Logistic regression provides simplicity and interpretability, support vector classifier excels in handling complex decision boundaries, and random forest harnesses the power of ensemble learning. By leveraging these classifiers, we've tailored our approach to the nuances of our news articles, achieving notable accuracy and performance. Our empirical results demonstrate that a thoughtful selection of models, aligned with the specific requirements of the task, can yield effective solutions in the realm of news categorization.

METHODOLOGY

A. ***Data Collection***

For this project a comprehensive dataset from kaggle comprising approximately 210,000 news headlines was curated from HuffPost over a decade, spanning from 2012 to 2022. The majority of the dataset, containing about 200,000 headlines, originates from the era when HuffPost's archival system was fully operational, between 2012 and May 2018. Below figure 1 shows the initial structure of data.


```
{
  "link": "https://www.huffpost.com/entry/covid-boosters-uptake-us_n_632d719ee4b087fae6feaac9",
  "headline": "Over 4 Million Americans Roll Up Sleeves For Omicron-Targeted COVID Boosters",
  "category": "U.S. NEWS",
  "short_description": "Health experts said it is too early to predict whether demand would match up with the 171 million doses of the new boosters the U.S. ordered for the fall.",
  "authors": "Carla K. Johnson, AP",
  "date": "2022-09-23"
}{
  "link": "https://www.huffpost.com/entry/american-airlines-passenger-banned-flight-attendant-punch-justice-department_n_632e25d3e4b0e247890329fe",
  "headline": "American Airlines Flyer Charged, Banned For Life After Punching Flight Attendant On Video",
  "category": "U.S. NEWS",
  "short_description": "He was subdued by passengers and crew when he fled to the back of the aircraft after the confrontation, according to the U.S. attorney's office in Los Angeles.",
  "authors": "Mary Papenfuss",
  "date": "2022-09-23"
}{
  "link": "https://www.huffpost.com/entry/funniest-tweets-cats-dogs-september-17-23_n_632de332e4b0695c1d81dc02",
  "headline": "23 Of The Funniest Tweets About Cats And Dogs This Week (Sept. 17-23)",
  "category": "COMEDY",
  "short_description": "\"Until you have a dog you don't understand what could be eaten.\"",
  "authors": "Elyse Wanshel",
  "date": "2022-09-23"
}{
  "link": "https://www.huffpost.com/entry/funniest-parenting-tweets_l_632d7d15e4b0d12b5403e479",
  "headline": "The Funniest Tweets From Parents This Week (Sept. 17-23)",
  "category": "PARENTING",
  "short_description": "\"Accidentally put grown-up toothpaste on my toddler\u2019s toothbrush and he screamed like I was cleaning his teeth with a Carolina Reaper dipped in Tabasco sauce.\"",
  "authors": "Caroline Bologna",
  "date": "2022-09-23"
}{
  "link": "https://www.huffpost.com/entry/amy-cooper-loses-discrimination-lawsuit-franklin-templeton_n_632c6463e4b09d8701bd227e",
  "headline": "Woman Who Called Cops On Black Bird-Watcher Loses Lawsuit Against Ex-Employer",
  "category": "U.S. NEWS",
  "short_description": "Amy Cooper accused investment firm Franklin Templeton of unfairly firing her and branding her a racist after video of the Central Park encounter went viral.",
  "authors": "Nina Golgowski",
  "date": "2022-09-22"
```


Fig. 1: Initial structure of data

B. ***Data Pre-processing***

The data pre-processing stage was crucial for optimizing the dataset for the machine learning project. It began with the identification and removal of empty values across the dataset to avoid biases and inaccuracies in model training. The 'short_description' column was particularly addressed by replacing empty entries with corresponding headlines and further converting any residual empty strings to NaN values before dropping them to preserve data completeness. To combat class imbalance, the dataset underwent an upsampling procedure, equalizing the representation across the top 15 categories, thereby preventing model bias towards more prevalent classes.

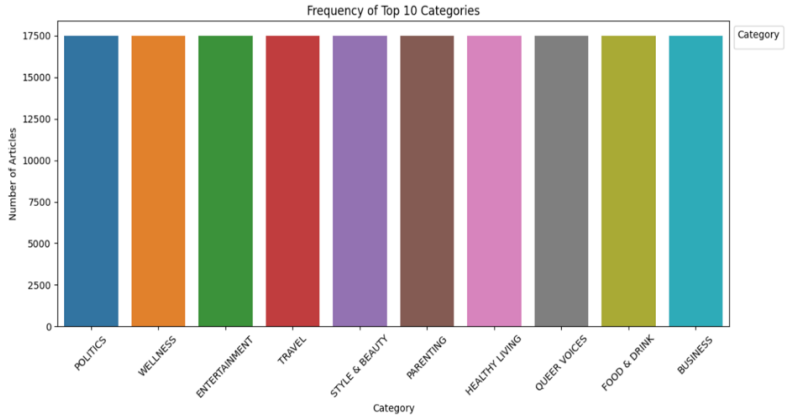

Fig. 2: Distribution of data after upsampling

Text normalization played a key role in this process, utilizing a SnowballStemmer for stemming and an augmented list of stopwords to refine the text data, eliminating noise and focusing on significant words. The text was further cleaned and standardized through various operations, such as converting to lowercase, removing HTML tags, hyperlinks, punctuation, newline characters, and numerically infused words. The pre-processing culminated with a sequential application of cleaning, stopwords removal, and stemming to prepare the dataset for the subsequent stages of model training and evaluation. This multi-faceted approach to text pre-processing was meticulously designed to refine the dataset into a form that is optimal for the subsequent application of NLP techniques. Below figure 3 is the final clean text after preprocessing.

	category	short_description	clean_text
63553	FOOD & DRINK	Don't just eat like a king, eat like the king.	dont eat like king eat like king
13002	BUSINESS	Truthfully, the highest credit score is not th...	truth highest credit score score shoot whole p...
7666	QUEER VOICES	“Well, maybe just a gorgeous facial hair momen...	“well mayb gorgeous facial hair moment like — ...
23356	WELLNESS	There's a place for Lycra. Spin class perhaps,...	there place lycra spin class perhap mayb run p...
76993	POLITICS	“Everything is off now, and Stormy is going to...	“everyth stormi go tell story” former porn act...
...	...	...	...
94515	HEALTHY LIVING	The research did not include mental health ser...	research includ mental health servic provid sc...
73423	ENTERTAINMENT	"You brought us so much joy."	brought us much joy
82871	QUEER VOICES	Sometime during the past fifteen years, I stop...	sometim past fifteen year stop oblivi instead ...
101576	STYLE & BEAUTY	Lawley has a point. Perhaps if we work to shif...	lawley point perhap work shift focus away bodi...
3438	WELLNESS	Great relationships are generally co-created b...	great relationship general cocreat peopl will ...

Fig. 3: Final cleaned text

C. ***Proposed Models***

Three distinct machine learning models were selected for the project aimed at categorizing news articles: Logistic Regression, Random Forest, and Support Vector Classifier. Logistic Regression was chosen for its simplicity, interpretability, and efficiency in multi-class problems, serving as a robust baseline model. The Random Forest model was included for its ability to handle the high dimensionality of text data, its resistance to overfitting, and its proficiency in capturing non-linear relationships, making it ideal for complex classification tasks. Support Vector Classifier was selected for its effectiveness in high-dimensional spaces and its capability to distinctly separate classes using the kernel trick, especially useful when dealing with clear margins of separation in data. Together, these models complement each other's strengths and offer a robust approach to text classification in news categorization.

D. ***Data Preparation***

In the course of this investigation, we meticulously enhanced the quality of our dataset through a comprehensive refinement process utilizing the TF-IDF technique. TF-IDF, recognized as a potent text analysis tool, played a pivotal role in evaluating the significance of words within our text corpus. This strategic refinement substantially elevated the overall quality of our dataset, facilitating an improved understanding for our machine learning models regarding the relative importance of each term. Our commitment to optimizing the performance of the TF-IDF algorithm led us to meticulously fine-tune its parameters. In this endeavor, we selectively excluded excessively common words that contributed minimal value, while retaining those that demonstrated sufficient frequency to justify inclusion. Notably, our parameterization extended to the evaluation of both single terms and bigram phrases, thereby enhancing the depth of our contextual analysis.

In addition to dataset refinement, we implemented a systematic division of the dataset into three segments, adhering to a balanced 60:20:20 ratio for training, validation, and testing purposes. This strategic partitioning proved to be of paramount importance in the meticulous tuning of hyperparameters governing our machine learning models. The allocation of distinct portions for training, validation, and testing facilitated a robust evaluation process, ensuring that our models were finely tuned and adept at generalizing to unseen data. Through these concerted efforts, we not only optimized the utility of the TF-IDF technique but also fostered a comprehensive approach to dataset management and model refinement, thereby enhancing the overall efficacy of our text analysis framework.

MODEL VALIDATION, EVALUATION AND RESULTS

In the domain of text classification, such as news categorization, it is critical to adopt metrics that reflect the model's ability to accurately predict the category of a new article. For this project, a combination of evaluation metrics was employed: accuracy, precision, recall, and the F1 score.

Accuracy: Accuracy serves as the most intuitive performance measure, providing a straightforward indication of a model's overall effectiveness. This metric was utilized as a primary measure due to its simplicity and clear interpretation, offering a quick snapshot of model performance. To address multi-class classification scenario, additional metrics were incorporated.

Precision: High precision relates to the low false positive rate, which in the context of news categorization, translates to a model's ability to reliably identify articles of a specific category.

Recall measures the model's capability to find all the relevant cases within a dataset. High recall indicates that the class is correctly recognized to a large extent. In terms of news categorization, recall is indicative of the model's ability to detect all articles belonging to a particular category, ensuring comprehensive coverage.

The F1 Score is particularly useful when the class distribution is uneven. The F1 Score is a better measure to use when seeking a balance between Precision and Recall and there is an uneven class distribution, as is often the case in news datasets.

Below figures 4, 5 and 6 show the classification report without hyperparameter tuning of the models Logistic regression, Random Forest and Support vector classifier respectively.

Accuracy: 0.8063142857142858
Classification Report:

	precision	recall	f1-score	support
BUSINESS	0.83	0.84	0.84	3457
ENTERTAINMENT	0.74	0.72	0.73	3546
FOOD & DRINK	0.85	0.90	0.87	3495
HEALTHY LIVING	0.81	0.73	0.77	3545
PARENTING	0.79	0.82	0.80	3417
POLITICS	0.79	0.78	0.79	3445
QUEER VOICES	0.89	0.80	0.84	3539
STYLE & BEAUTY	0.86	0.85	0.85	3539
TRAVEL	0.84	0.84	0.84	3532
WELLNESS	0.69	0.78	0.73	3485
accuracy			0.81	35000
macro avg	0.81	0.81	0.81	35000
weighted avg	0.81	0.81	0.81	35000

Fig. 4: Classification report of Logistic Regression without HT

Random Forest Model Accuracy: 0.8984857142857143

	precision	recall	f1-score	support
BUSINESS	0.94	0.95	0.95	3457
ENTERTAINMENT	0.75	0.86	0.80	3546
FOOD & DRINK	0.92	0.96	0.94	3495
HEALTHY LIVING	0.89	0.90	0.90	3545
PARENTING	0.95	0.91	0.93	3417
POLITICS	0.86	0.86	0.86	3445
QUEER VOICES	0.94	0.93	0.94	3539
STYLE & BEAUTY	0.94	0.90	0.92	3539
TRAVEL	0.94	0.90	0.92	3532
WELLNESS	0.88	0.80	0.84	3485
accuracy			0.90	35000
macro avg	0.90	0.90	0.90	35000
weighted avg	0.90	0.90	0.90	35000

Fig. 5: Classification report of Random Forest without HT

SVC Model Accuracy: 0.8544

	precision	recall	f1-score	support
BUSINESS	0.89	0.89	0.89	3457
ENTERTAINMENT	0.75	0.80	0.77	3546
FOOD & DRINK	0.90	0.92	0.91	3495
HEALTHY LIVING	0.87	0.79	0.83	3545
PARENTING	0.86	0.87	0.86	3417
POLITICS	0.83	0.82	0.83	3445
QUEER VOICES	0.93	0.85	0.89	3539
STYLE & BEAUTY	0.90	0.89	0.90	3539
TRAVEL	0.90	0.88	0.89	3532
WELLNESS	0.75	0.83	0.79	3485
accuracy			0.85	35000
macro avg	0.86	0.85	0.86	35000
weighted avg	0.86	0.85	0.86	35000

Fig. 6: Classification report of Support Vector Classifier without HT

For Logistic Regression, the model achieved an accuracy of about 80.63%, with precision, recall, and F1-scores mostly above 0.7 across various classes, indicating a reliable predictive performance. The Random Forest Classifier showed a significant improvement with an accuracy of approximately 89.85%. The classification metrics were high across the board, reflecting robust classification ability. The Support Vector Classifier's accuracy was noted to be 85.44%, with individual class scores for precision, recall, and F1-score generally high, suggesting strong classification but with some variation between classes.

Below figures 7 and 8 show the classification report with hyperparameter tuning of the models Logistic regression, Random Forest respectively

Best Hyperparameters: OrderedDict([('C', 9.971154848447195), ('penalty', 'l1'), ('solver', 'liblinear')])
Accuracy: 0.8643428571428572
Classification Report:

	precision	recall	f1-score	support
BUSINESS	0.87	0.93	0.90	3457
ENTERTAINMENT	0.81	0.78	0.79	3546
FOOD & DRINK	0.90	0.93	0.92	3495
HEALTHY LIVING	0.84	0.85	0.84	3545
PARENTING	0.86	0.89	0.88	3417
POLITICS	0.86	0.81	0.84	3445
QUEER VOICES	0.90	0.89	0.90	3539
STYLE & BEAUTY	0.90	0.91	0.90	3539
TRAVEL	0.89	0.89	0.89	3532
WELLNESS	0.80	0.77	0.78	3485
accuracy			0.86	35000
macro avg	0.86	0.86	0.86	35000
weighted avg	0.86	0.86	0.86	35000

Fig. 7: Classification report of Logistic Regression with HT

Best Hyperparameters for Random Forest:
OrderedDict([('max_depth', 20), ('max_features', 'log2'), ('min_samples_leaf', 1), ('min_samples_split', 20), ('n_estimators', 1000)])
Accuracy for Random Forest: 0.6321142857142857
Classification Report for Random Forest:

	precision	recall	f1-score	support
BUSINESS	0.84	0.55	0.67	3457
ENTERTAINMENT	0.64	0.50	0.56	3546
FOOD & DRINK	0.47	0.90	0.61	3495
HEALTHY LIVING	0.83	0.35	0.49	3545
PARENTING	0.57	0.80	0.66	3417
POLITICS	0.48	0.82	0.60	3445
QUEER VOICES	0.90	0.52	0.65	3539
STYLE & BEAUTY	0.80	0.69	0.74	3539
TRAVEL	0.86	0.59	0.70	3532
WELLNESS	0.61	0.63	0.62	3485
accuracy			0.63	35000
macro avg	0.70	0.63	0.63	35000
weighted avg	0.70	0.63	0.63	35000

Fig. 8: Classification report of Random Forest with HT

Bayesian optimization tuned models showed varied results. For Logistic Regression, the accuracy reached around 86.43%, and for the Random Forest, it was around 63.21%. The optimization had mixed impacts, with Logistic Regression showing relatively consistent performance across classes, while the Random Forest model demonstrated a drop in accuracy and more pronounced misclassifications.

In conclusion, these models demonstrate the intricacies of model selection and the importance of tuning in achieving higher accuracy in machine learning tasks. Each model has strengths and weaknesses that are revealed through these performance metrics, emphasizing the necessity of careful model evaluation and selection in practice.

DISCUSSIONS AND FUTURE IMPROVEMENTS

Upon reflection of the project's outcomes and the insights garnered throughout the process of automated news categorization, there emerges a platform for discussion on the achievements and areas ripe for future improvements. The endeavor of categorizing news using machine learning techniques has yielded promising results, demonstrating the potential of Logistic Regression, Random Forest, and Support Vector Classifier models in addressing the complex task of text classification. However, the landscape of NLP and machine learning is one of continual advancement, and the pursuit of enhanced performance and sophistication in models remains an ongoing commitment. A critical observation from the project is the interplay between the chosen evaluation metrics and the models' performance. While the models achieved respectable accuracy, precision, recall, and F1 scores, there is an acknowledgment that these metrics could be further optimized. The balance between precision and recall is a delicate one, often requiring a nuanced approach to model training and parameter tuning. Future efforts could explore the use of alternative or additional metrics, such as the area under the receiver operating characteristic curve (AUC-ROC) for each category, to gain deeper insights into model performance across various thresholds. The scalability and adaptability of the models to the ever-evolving nature of news content present another avenue for development. The project could benefit from incorporating a dynamic training regime where the models are regularly updated with new data. This would not only improve the models' responsiveness to the latest news trends and linguistic usage but also enhance their ability to generalize across different periods and styles of news writing. Advancements in deep learning offer substantial promise for the future of news categorization. Models such as transformer-based architectures, including BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer), have revolutionized the field of NLP. Integrating such models could significantly bolster the system's accuracy and efficiency. Their ability to understand context and capture subtleties in language could prove invaluable in differentiating between closely related news categories and managing the nuances of sarcasm, irony, and humor often found in news headlines.

A further consideration for future work is the exploration of unsupervised and semi-supervised learning techniques. These methods could be particularly beneficial in addressing the scarcity of labeled data for certain categories or in leveraging the large volumes of unlabeled text data available online. Unsupervised topic modeling or clustering could reveal underlying patterns and themes in the data that supervised models might overlook. In conclusion, while the project has established a strong foundation for automated news categorization, there is a multitude of pathways to enhance and refine the system. Future improvements could focus on incorporating cutting-edge deep learning models, expanding the dataset, employing a continuous learning approach, and experimenting with unsupervised learning techniques. These efforts would not only improve the system's performance but also ensure its relevance and efficacy in the dynamic field of news categorization.

REFERENCES

[1] S. Usmani and J. A. Shamsi, "News Headlines Categorization Scheme for Unlabelled Data," 2020 International Conference on Emerging Trends in Smart Technologies (ICETST), Karachi, Pakistan, 2020, pp. 1-6, doi: 10.1109/ICETST49965.2020.9080726.

[2] A. Rizaldy and H. A. Santoso, "Performance improvement of Support Vector Machine (SVM) With information gain on categorization of Indonesian news documents," 2017 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2017, pp. 227-232, doi: 10.1109/ISEMANTIC.2017.8251874.

[3] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev and A. Alizade, "Empirical Study of Online News Classification Using Machine Learning Approaches," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, 2018, pp. 1-6, doi: 10.1109/ICAICT.2018.8747012.

[4] Y. Liu, M. He, M. Shi and S. Jeon, "A Novel Model Combining Transformer and Bi-LSTM for News Categorization," in IEEE Transactions on Computational Social Systems, doi: 10.1109/TCSS.2022.3223621.

[5] Misra, Rishabh. "News Category Dataset." arXiv preprint arXiv:2209.11429 (2022).

[6] Misra, Rishabh and Jigyasa Grover. "Sculpting Data for ML: The first act of Machine Learning." ISBN 9798585463570 (2021).