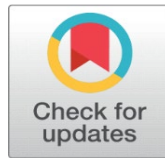


NEWS WEB APPLICATION EVALUATION USING MACHINE LEARNING

Dr. Sunil Rathod ¹, Dr. Vikas Nandgaoknar ¹, Rutika Chougale ¹

¹Department of Computer Engineering, Indira College of Engineering and Management, Pune, India



DOI

[10.29121/shodhkosh.v4.i2.2023.5726](https://doi.org/10.29121/shodhkosh.v4.i2.2023.5726)

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Copyright: © 2023 The Author(s). This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

With the license CC-BY, authors retain the copyright, allowing anyone to download, reuse, re-print, modify, distribute, and/or copy their contribution. The work must be properly attributed to its author.



ABSTRACT

In our contemporary digital landscape, news consumption has transitioned largely to online platforms, shaping public opinion and influencing societal discourse. However, this shift has also led to the proliferation of misinformation and fake news, which can have profound consequences on public perception and decision-making processes. Moreover, the rise of social media has accelerated the dissemination of news, amplifying the impact of false information.

In response to these challenges, this project focuses on developing a robust framework for evaluating news websites using advanced techniques in Artificial Intelligence (AI), Natural Language Processing (NLP), and Machine Learning (ML). The primary objective is to empower users with the ability to discern between credible and unreliable sources of information by performing binary classification of news articles.

Through the utilization of sophisticated algorithms and datasets, our system aims to analyse the content of online news articles and assess their authenticity. By leveraging AI and ML models, users will be equipped with tools to identify potentially misleading or fabricated news stories, thereby promoting critical thinking and informed decision-making.

Key features of the proposed system include the classification of news articles as either authentic or fake, as well as an evaluation of the credibility of the websites publishing the news. By harnessing the power of AI-driven analysis, this project endeavours to mitigate the spread of misinformation and enhance trust in online news sources.

Keywords: Graphical Representation, Flask Framework, Multiple Website Evaluation, Linear Regressor

1. INTRODUCTION

In the digital age, the internet serves as a primary source of information for individuals worldwide. However, in the vast array of websites and content available online, ensuring the authenticity and reliability of news sources has become increasingly challenging. With the proliferation of fake news websites and misinformation campaigns, there is a pressing need for effective tools to evaluate the credibility of online news sources.

The project "News website evaluation using opinion mining" addresses this need by leveraging machine learning techniques to analyse and evaluate the authenticity of news websites. By harnessing the power of sentiment analysis and classification algorithms, the system aims to provide users with valuable insights into the trustworthiness and credibility of news sources.

Through a combination of static and dynamic analysis methodologies, the project offers a comprehensive approach to news website evaluation. The static component utilizes machine learning classifiers to assess the inherent characteristics of news websites, while the dynamic aspect involves real-time analysis of content to determine the probability of truthfulness.

Key technologies such as Python, Flask, Sci-Kit Learn, and web scraping tools are employed to develop a user-friendly web interface that allows users to input URLs and receive accurate evaluations of news website authenticity. Additionally, the system incorporates features for user feedback and interaction, enabling continuous improvement and refinement.

By empowering users with the ability to make informed decisions about the reliability of news sources, this project contributes to combating misinformation and promoting media literacy in the digital era. Through its innovative approach to news website evaluation, the project seeks to enhance trust in online information and foster a more informed society.

2. LITERATURE SURVEY

[1] Priyanka Rathore, Dr. Anurag Jain, and Chetan Agrawal present a comprehensive analysis of methodologies for predicting the popularity of online news articles in their paper. They compare and evaluate different predictive models, taking into account various parameters to discern their effectiveness. Additionally, the authors propose enhancements and refinements to existing methodologies, culminating in the development of a robust online news popularity prediction system.

The research landscape surrounding online news popularity prediction is multifaceted and dynamic. By synthesizing insights from prior studies and leveraging advanced machine learning techniques, Rathore et al. contribute to the evolving discourse on this topic. Their proposed methodology serves as a valuable framework for stakeholders seeking to harness predictive analytics for informed decision-making in the digital news domain.

[2] P. Keerthana, B. Meghana, P. Akshaya, 'Website Evaluation Using Opinion Mining', 2021. evaluation holds significant promise for businesses seeking to understand and cater to user preferences in the digital domain. By harnessing the power of user-generated content, businesses can gain actionable insights into consumer behavior, preferences, and sentiments, thereby improving customer satisfaction and loyalty. However, challenges such as data privacy, bias detection, and algorithmic transparency must be addressed to ensure the ethical and responsible use of opinion mining technologies. Additionally, future research should focus on advancing methodological frameworks, enhancing algorithmic accuracy, and exploring novel applications of opinion mining in website evaluation.

This feedback serves as invaluable data for businesses involved in website promotion, enabling them to refine designs, personalize user experiences, and gain deeper insights into website performance. Moreover, these reviews play a crucial role for individuals seeking recommendations or references for websites. Opinion mining technology, applied in this context, facilitates the comparison of websites across different brands based on user ratings. Such an approach aids consumers in making informed decisions aligned with their preferences. This literature review explores the utilization of opinion mining techniques for evaluating website performance and standards, with the primary objective of providing users with a comprehensive rating system.

[3] In 'A review on sentiment analysis from social media platforms' paper, Margarita Rodríguez-Ibáñez, Antonio Casañez-Ventura, Félix Castejón-Mateos, Pedro-Manuel Cuenca-Jiménez propose a comprehensive review of sentiment analysis in social networks, delving into both established methodologies and emerging trends. We aim to not only scrutinize existing methods from an academic standpoint but also to explore novel dimensions such as temporal dynamics, causal relationships, and industrial applications. Our examination extends to diverse domains where sentiment analysis is applicable, with a particular focus on its implications for stock market dynamics, political discourse, and addressing cyberbullying in educational environments.

Mykhailo Granik et. al., their paper shows a basic methodology for counterfeit news location utilizing guileless Bayes classifier. This approach was carried out as a product framework and tried against an informational index of Facebook news posts. They were gathered from three enormous Facebook pages each from the right and from the left, as well as three huge standard political news pages (Politico, CNN, ABC News). They accomplished order exactness of around 74%. Order exactness for counterfeit news is somewhat more terrible. This might be brought about by the skewness of the dataset: just 4.9% of it is phony information. Himank Gupta et. al. gave a structure in light of various AI approach that arrangements with different issues including exactness deficiency, delay (BotMaker) and high handling time to deal with large number of tweets in 1 sec. They, first and foremost, have gathered 400,000 tweets from HSpam14 dataset.

Then they further describe the 150,000 spam tweets and 250,000 non-spam tweets. They additionally inferred a few lightweight elements alongside the Main 30 words that are giving most elevated data gain from Pack of Words model. They had the option to accomplish a precision of 91.65% and outperformed the current arrangement by roughly 18%. Marco L. Della Vedova et. al. first proposed a clever ML counterfeit news identification strategy which, by joining news content and social setting highlights, beats existing techniques in the writing, expanding its precision up to 78.8%.

















Second, they carried out their strategy inside a Facebook Courier Chabot and approve it with a genuine application, getting a phony news recognition exactness of 81.7%.

Their objective was to characterize a news thing as dependable or counterfeit; they previously portrayed the datasets they utilized for their test, then introduced the substance-based approach they executed and the technique they proposed to join it with a social-based approach accessible in the writing. The subsequent dataset is made out of 15,500 posts, coming from 32 pages (14 intrigue pages, 18 logical pages), with more than 2, 300, 00 preferences by 900,000+ clients. 8,923 (57.6%) posts are scams and 6,577 (42.4%) are non-lies.

Table-1 Literature Survey

Sr. No.	Title of Paper	Year	Author	Key Points	Limitations	Gap identified
1	A Comprehensive Review on Online News Popularity Prediction using Machine Learning Approach	2019	Anurag Jain, Chetan Agrawal	<ul style="list-style-type: none"> News articles online news popularity Popularity prediction 	1. Multilingual . Content	Ambiguity
2	Website Evaluation Using Opinion Mining	2021	P Keerthana, B Meghana, P Akshaya	<ul style="list-style-type: none"> Social networking Websites Internet sites Opinion mining Rating 	1.Lack of human understanding. 2.Context understanding	Negation handling
3	A review on sentiment analysis from social media platforms	2023	Margarita Rodríguez-Ibanez ^{a,*} , Antonio Casanez-Ventura ^b , Félix Castejon-Mateos ^b , Pedro-Manuel Cuenca-Jiménez ^b	<ul style="list-style-type: none"> Sentiment analysis Social media Twitter sentiment analysis 	1. Limited to text 2. By evolving language and trends	1.Context Sensitivity.

Table 2 Taxonomy Chart, Comparison of existing system with proposed system

Parameters System ↓	Multiple Website Evaluation	Detailed Design report	Predict the Changes	Ease of access
A Comprehensive Review on Online News Popularity Prediction using Machine Learning Approach				
Website Evaluation Using Opinion Mining				
A review on sentiment analysis from social media platforms				
Proposed System				

The Table-2 shows the comparison of the existing systems and the proposed system based on the important features used as the evaluation parameters.

In [1] the system does not support the content evaluation of multiple sites and also the system is not easy to access.

In [2] the system does not provide the detail design report based on which the popularity of the website can be predicted.

In [3] the system lacks in the doing the evaluation of multiple websites and also it is not able to predict the changes.

In our proposed system, the limitation of all the existing systems under consideration are overcome using machine learning model with the proper design of the dataset.

3. PROPOSED SYSTEM ARCHITECTURE

Static Search

The architecture of Static part of fake news detection system is quite simple and is done keeping in mind the basic machine learning process flow. The system design is shown below and self-explanatory. The main processes in the design are:

1) User Interface Layer:

- **Web Interface:** Users interact with the system through a web-based interface developed using HTML, CSS, and JavaScript.
- **Input Forms:** Users input the URL of the news website they want to evaluate and provide keywords or text for dynamic truth probability analysis.
- **Feedback Mechanism:** Optionally, users may provide feedback on the evaluation process.

2) Application Layer:

- **Flask Framework:** Flask serves as the web application framework for routing user requests, handling sessions, and rendering templates.
- **Request Handling:** Incoming requests are processed by Flask routes, which trigger appropriate actions within the application.
- **Data Processing:** The application processes user input, conducts sentiment analysis, and retrieves online information for truth probability assessment.

3) Model Layer:

- **Machine Learning Model:** This layer includes the trained machine learning model for opinion mining and authenticity verification.
- **Sci-Kit Learn Integration:** The model, developed using Sci-Kit Learn, is integrated into the system to perform sentiment analysis and classification tasks.

4) Data Layer:

- **SQLAlchemy ORM:** SQL Alchemy is used to interact with the database, storing user information, feedback, and other relevant data.
- **Database:** The system may utilize a relational database to store user data and model training information.

5) External Services:

- **Web Scraping Tools:** Requests and BeautifulSoup are employed to scrape online content for dynamic truth probability analysis.
- **Third-Party APIs:** External APIs may be utilized for additional functionalities, such as social media impact analysis or URL validation.

6) Integration and Communication:

- **Data Flow:** Data flows between the various layers of the system, with Flask handling communication between the user interface, application logic, and external services.
- **Model Integration:** The machine learning model is integrated into the application layer, where it processes user input and provides evaluation results.
- **API Integration:** External APIs or web scraping tools are integrated into the application to gather supplementary information for evaluation.

7) Security Layer:

- **Encryption:** User data, including passwords and session information, may be encrypted using bcrypt to enhance security.
- **Input Validation:** Input validation techniques are implemented to prevent malicious input and protect against potential security threats.

4. PROPOSED SYSTEM

The proposed system architecture is shown in Figure-1.

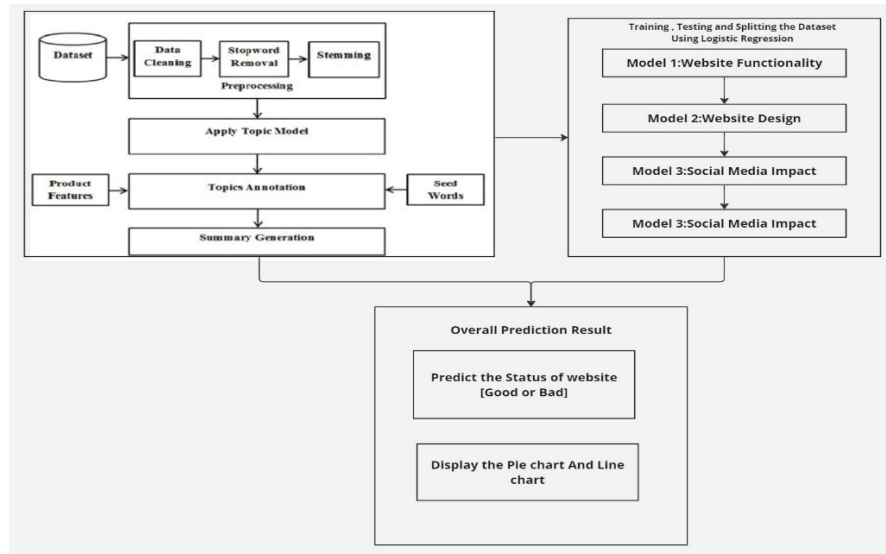


Figure 1 Proposed System Architecture

User Authentication and Dashboard: The system begins with a user authentication mechanism, allowing registered users to access the evaluation features. Upon successful login, users are directed to a personalized dashboard interface.

Website Input and Preview: Users are provided with an input field where they can enter the URL of the news website they wish to evaluate. Upon submission, the system retrieves and displays a preview of the website within a designated window for reference.

Analysis and Evaluation: After the website preview is displayed, users can initiate the analysis process by clicking on a designated button. The system then performs opinion mining on various aspects of the website, including functionality, social media impact, and design quality.

Input Fields for Evaluation Criteria: Following the initial analysis, users are directed to a subsequent page featuring input fields for each evaluation criterion. These criteria include website functionality, social media impact, and website design. Users can provide their opinions or ratings based on predefined scales or criteria.

Model Training and Testing: The system utilizes datasets to train and test the underlying opinion mining model. This model is continuously refined to enhance accuracy and relevance in evaluating news websites.

Results Presentation: Upon completion of the evaluation process, users are presented with a final results page. Here, the system provides a comprehensive assessment of the website's functionality, social media impact, and design quality. Results are categorized as either favourable or unfavourable based on predefined thresholds or criteria.

Visual Representation: To enhance user understanding and facilitate comparison, the system includes visual aids such as pie charts depicting website content distribution and line charts illustrating trends in functionality, social media impact, and design over time.

User Feedback Mechanism: Finally, the system incorporates a feedback mechanism allowing users to provide their feedback on the evaluation process or suggest improvements for future iterations.

5. IMPLEMENTATION AND EXPERIMENTAL RESULTS

The point of this framework is to improve ideas about the plan of site or web application by examining clients' perspective and opinions.

The framework will actually want to distinguish the regions where clients are disappointed with the site. With the utilization of this data, we can roll out specific improvements in the plan of the application that will upgrade client fulfilment and commitment.

It will likewise assist with recognizing issues connected with content quality, pertinence, and precision of the web application.

- Further develop the general client experience by recognizing and addressing client concerns connected with web architecture, content quality, and convenience.
- Guarantee that news stories and content line up with client inclinations and interests, improving substance importance and commitment.
- Use client criticism and opinion examination to drive information driven decision-production for persistent upgrades in web architecture and content curation.
- Enhance site/application execution by resolving specialized issues and guaranteeing a consistent client experience

System improvement is likewise taken into consideration as a system subsidized with the aid of using engineering techniques. We have attempted to incorporate & expand new debris for our training debris were observed now no longer all through the coding however additionally all through the analysis, layout phases & in documentation.

Website evaluated the use of opinion mining assignment is taken into consideration as an enlargement of commercial enterprise relations. It contributes lots with the aid of using imparting quick & rapid offerings of sending files letters (formal & casual both) to commercial enterprise because it allows any commercial enterprise to flourish Following change or improvements may be finished in the system.

- 1) More than one organization may be incorporated via this software.
- 2) Web offerings may be used to realize the genuine transport popularity of packets.
- 3) Clients can test the repacked transport popularity online.
- 4) Distributed database technique in location of centralized technique.

1) Methodology

This paper outlines a system developed in three key components. The first part is static and revolves around machine learning classifiers. Initially, four different classifiers were studied and trained, with the optimal classifier selected for final implementation. The second part is dynamic and involves taking keywords or text input from the user and searching online to determine the truth probability of the news. Lastly, the third part focuses on verifying the authenticity of URLs inputted by the user.

Python, along with its Sci-kit libraries, serves as the primary technology stack for this project. Python offers a vast array of libraries and extensions, making it conducive to machine learning tasks. The Sci-Kit Learn library stands out as a comprehensive resource for various machine learning algorithms, enabling straightforward evaluation and implementation.

For web-based deployment of the model, Flask is utilized, providing a robust framework for creating web applications. The client-side implementation is facilitated through HTML, CSS, and JavaScript, ensuring an interactive user experience. Additionally, BeautifulSoup (bs4) and requests are employed for web scraping tasks, enabling dynamic retrieval of online information.

This approach leverages the strengths of Python's ecosystem and web development frameworks like Flask to create a versatile and effective system for evaluating news authenticity and URL validity.

A statistical model typically used to model a binary dependent variable with the help of logistic function. Another name for the logistic function is a sigmoid function and is given by:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

This function assists the logistic regression model to squeeze the values from $(-k, k)$ to $(0, 1)$. Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

In the execution of the venture "News site assessment utilizing assessment mining," the emphasis was on fostering an exhaustive assessment framework. The cycle starts with an easy-to-understand login and enrolment connection point to work with client cooperation. Upon fruitful confirmation, clients are coordinated to a powerful dashboard where they can enter the URL of the news site they wish to assess.

After entering the URL, the site is reviewed inside an assigned window for simplicity of reference. In this manner, after setting off the examination cycle by tapping the assigned button, clients are coordinated to a resulting page highlighting input fields for assessing site usefulness, online entertainment effect, and web composition.

To play out these assessments, datasets were utilized to prepare and test the hidden assessment mining model. This model successfully investigates the site against predefined standards, considering an educated assessment.

The summit of this assessment interaction is introduced to the client on an eventual outcomes page. Here, clients are furnished with a far-reaching evaluation of the site's usefulness, virtual entertainment effect, and plan quality, sorted as either positive or troublesome. Also, visual guides, for example, pie diagrams portraying site content dispersion and line outlines showing the site's usefulness, web-based entertainment effect, and configuration patterns are incorporated to upgrade understanding and work with correlation.

2) Algorithms and Dataset Preparation

The only algorithm used in the implementation of the system is linear regression which unexpectedly has given best results. The datasets were created on the basis of the features selected for the evaluation in each of the cases of:

- Website Contents
- Website Design
- Website Functionality
- Social Media Impact of the news of Website

In case of Website Contents, it has been observed that the news that are popular are generally having more positive words than the negative words. Based on this binary class of positive and negative contents, the standard dataset from Kaggle was taken for the evaluation and traditional but powerful technique of Natural Language Processing (NLP) has been used to segregate the contents into positive words in the news article and the negative word in that article.

The design of the website has been evaluated based on layout, typography, colour scheme and responsive design. The dataset was prepared by observing the website of the given URL.

The functionality of any given website has been critically studied using parameters like load time, personalization, multimedia support and accessibility of the site.

The social media impact of given URL has been evaluated on the criteria of how many times the URL has been shared, how many likes and comments the URL has got and finally how many clicks or hits are there for the given URL.

The data set for Website Design, Functionality and the social media impact has been done by manual observation of some prominent news websites and the linear regressor classifier of machine learning was used to do the prediction. As the linear regressor classifier gives good result no other classifier was tried and tested for the evaluation of the given URL based on the already discussed parameter.

The overall prediction was done by comparing each of the evaluation parameter and appropriate suggestion were given to the owner of the website via mail and also shown to it on the dashboard of the system user.

The accuracy of the linear regressor model along with the confusion matrix is shown in Figure-2.

```
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
* Restarting with stat

ACCURACY SCORE of Web Design: 1.0
[[26  0]
 [ 0 1]]

ACCURACY SCORE of Web Functionality: 0.9761904761904762
[[16  1]
 [ 0 25]]

ACCURACY SCORE of Social Media: 0.96
[[18  0]
 [ 1 14]]

ACCURACY SCORE Overall Website: 1.0
[[1 0]
 [0 1]]
* Debugger is active!
* Debugger PIN: 118-714-237
```

Figure 2 Accuracy Score and the Confusion Matrix for each functionality of given URL

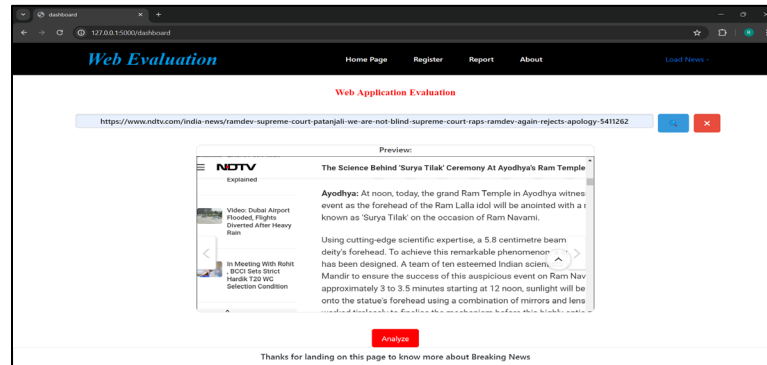


Figure 3 Home Page of Web Site Evaluation

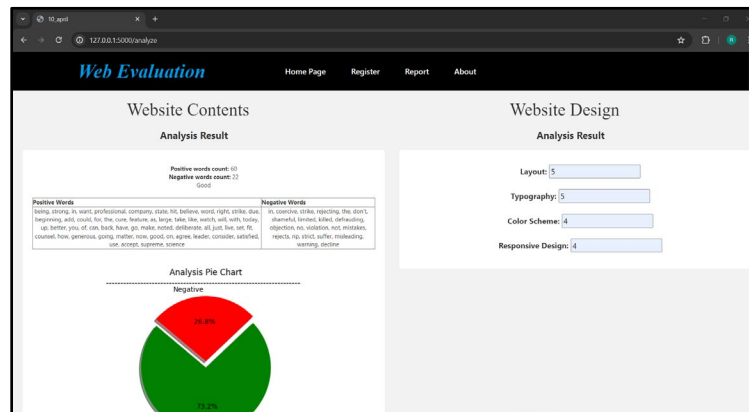


Figure 4 Analysis Page (1)

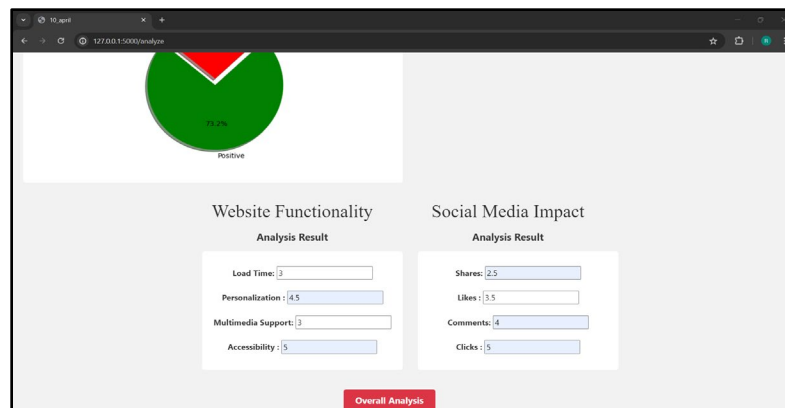


Figure 5 Analysis Page (2)

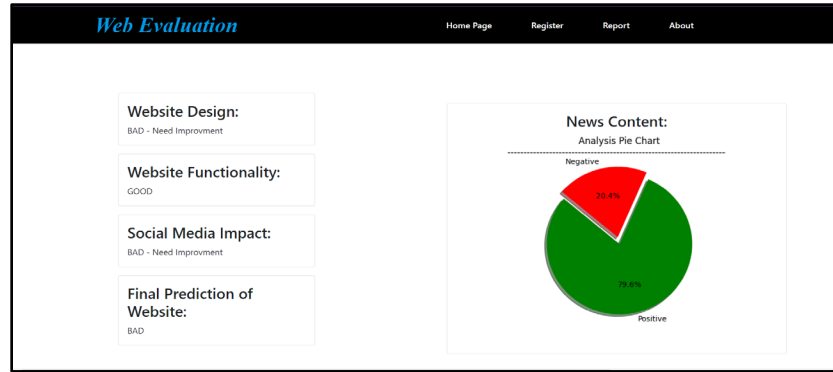


Figure 6 Overall Web Site Evaluation

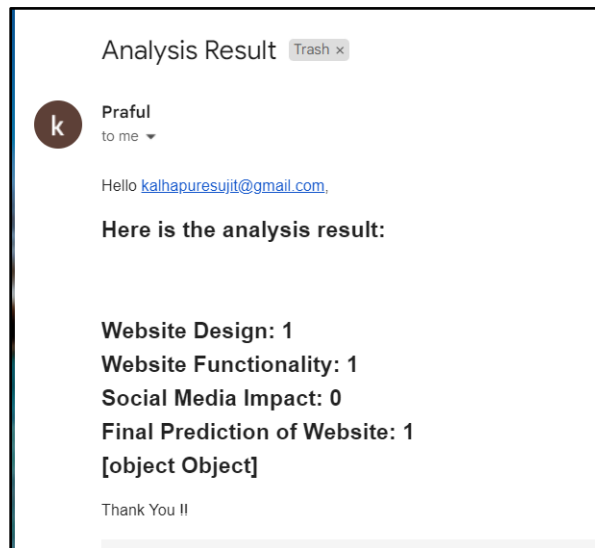


Figure 7 Email of the result Web Site Evaluation

6. CONCLUSION

In conclusion, the project "News website evaluation using opinion mining" has successfully demonstrated the effectiveness of employing sentiment analysis techniques to assess various aspects of news websites. Through the development of a user-friendly interface encompassing login, registration, and a dynamic dashboard, users are provided with a seamless experience for inputting website URLs and accessing evaluation results.

By utilizing datasets to train and test the opinion mining model, the project ensures robustness and accuracy in its evaluations of website functionality, social media impact, and design quality. The final results page offers users a clear and comprehensive overview of the evaluated website, presenting findings in a user-friendly manner through categorization and visual aids such as pie charts and line graphs.

Overall, the project underscores the potential of opinion mining techniques in providing valuable insights into the quality and effectiveness of news websites. Moving forward, further enhancements could be made to refine the evaluation process and expand the scope to encompass additional parameters, thereby contributing to the ongoing evolution of web evaluation methodologies.

Our system gives the customer a better understanding about websites. It prevents the customer from being a victim of fraud. It will give them a detailed survey about the service or customer feedback regarding those websites. Our system could be a critic for the websites hence driving them towards improvement on their drawbacks. Based on data availability our system can be expanded to different sectors. It is also open to customers' views based on their frequent experience with those websites.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

- Priyanka Rathord Dr. Anurag Jain Chetan Agrawal, 'A Comprehensive Review On Online News Popularity Prediction Using Machine Learning Approach', 2019.
- P. Keerthana, B. Meghana, P. Akshaya, 'Website Evaluation Using Opinion Mining', 2021.
- Margarita Rodríguez-Ibanez , Antonio Casanez-Ventura , F'elix Castejon-Mateos, Pedro-Manuel Cuenca-Jim'Enez 'A Review On Sentiment Analysis From Social Media Platforms', 2023
- <http://www.cs.virginia.edu/~up3f/cs4750/supplement/db-setup-xampp.html>
- Barbieri, Camacho-Collados, Espinosa Anke, & Neves, L. Tweet Eval: Unified Benchmark And Comparative Evaluation For Tweet Classification. In Findings Of The Association For Computational Linguistics: EMNLP 2020 Pp. 1644–1650
- Elena Hensinger, Ilias Flaounas, Nello Cristianini, "Modelling And Predicting News Popularity" Springer, Pattern Anal Applic, 2013, Pp. 623–635.
- Arapakis, Ioannis, B. Barla Cambazoglu, And Mounia Lalmas. "On The Feasibility Of Predicting News Popularity At Cold Start." International Conference On Social Informatics. Springer International Publishing, 2014.
- Al-Mutairi, Hanadi Muqbil, And Mohammad Badruddin Khan. "Predicting The Popularity Of Trending Arabic Wikipedia Articles Based On External Stimulants Using Data/Text Mining Techniques." Cloud Computing (ICCC), 2015 International Conference On. IEEE, 2015.
- Tatar Alexandru, Et Al. "Predicting The Popularity Of Online Articles Based On User Comments." Proceedings Of The International Conference On Web Intelligence, Mining And Semantics. ACM, 2011.
- Castillo, Carlos, Et Al. "Characterizing The Life Cycle Of Online News Stories Using Social Media Reactions." Proceedings Of The 17th ACM Conference On Computer Supported Cooperative Work & Social Computing. ACM, 2014.