JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

FAKE NEWS DETECTION SYSTEM

USING LOGISTIC REGRESSION, NLP &FEATURE ENGINEERING

SAMIRAN DEORE[1], KAUSHAL KOTKAR[2], SHARAD MANANI[3], RIDDHI TAK[4], PROF. SNEHA SHINGARE[5]

Cyber Security Engineering Department SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE , CHEMBUR, Univeristy of Mumbai

Abstract: This study proposes a Fake News Detection System (FNDS) utilizing logistic regression, natural language processing (NLP), and feature engineering. The system addresses the growing challenge of identifying fake news amidst vast amounts of online content. By leveraging logistic regression for classification, NLP for text processing, and feature engineering for informative feature extraction, FNDS achieves accurate detection. Experimental results demonstrate the effectiveness of this approach, showcasing its potential for combating misinformation in real-world settings.

I. INTRODUCTION

The proliferation of misinformation and fake news in online platforms has emerged as a critical societal challenge, influencing public opinion, political discourse, and social cohesion. In response to this phenomenon, the development of robust fake news detection systems has become imperative. This introduction provides a brief overview of a Fake News Detection System (FNDS) that harnesses the power of logistic regression, natural language processing (NLP), and feature engineering to combat the spread of misinformation In recent years, the advent of social media and digital communication platforms has facilitated the rapid dissemination of information, both genuine and fabricated. This unrestricted flow of content has created fertile ground for the propagation of fake news, which often masquerades as legitimate reporting, misleading unsuspecting readers Detecting such deceptive content manually is a daunting task, given the sheer volume of information circulating online.To address this challenge, researchers and data scientists have turned to machine learning and NLP techniques to develop automated fake news detection systems. These systems leverage advanced algorithms to analyze the linguistic features and contextual cues present in textual content, distinguishing between credible and deceptive information. Among the myriad of approaches, logistic regression has emerged as a popular choice for its simplicity, interpretability, and efficacy in binary classification tasks .In conjunction with logistic regression, NLP techniques play a pivotal role in preprocessing textual data, extracting meaningful features, and uncovering linguistic patterns indicative of fake news. By tokenizing text, removing stop words, and performing syntactic analysis, NLP enables the system to discern subtle linguistic nuances that may betray the authenticity of a news article. Moreover, feature engineering augments this process by crafting informative features from raw text, enriching the model's understanding of the underlying data.The integration of logistic regression, NLP, and feature engineering in a unified fake news detection framework represents a promising approach to addressing the challenges posed by misinformation. By leveraging the strengths of each component, the FNDS strives to achieve high accuracy, robustness, and scalability in identifying deceptive content across diverse online platforms.In summary, the FNDS embodies a multidisciplinary approach to combating the proliferation of fake news, drawing upon insights from machine learning, natural language processing, and data engineering. Through the synergistic integration of logistic regression, NLP, and feature engineering, the system aims to bolster the resilience of online communities against the pernicious effects of misinformation, fostering a more informed and discerning public discourse.

III. RELATED WORK

Numerous studies have delved into the realm of fake news detection, employing a combination of logistic regression, natural language processing (NLP), and feature engineering techniques.

- **Wang et al. (2017)**: Utilized logistic regression alongside NLP for discerning fake from real news articles. They extracted linguistic features like n-grams and sentiment analysis scores, showcasing the effectiveness of logistic regression in this domain.
- **Rubin et al. (2016)**: Introduced a framework employing logistic regression as the core classification algorithm, while leveraging NLP for text pre-processing and feature engineering. Their work emphasized the significance of feature engineering in capturing linguistic cues indicative of fake news.

- **Gupta et al. (2018)**: Explored the role of feature engineering in enhancing fake news detection models. By devising novel features based on linguistic patterns and readability metrics, they augmented the discriminatory power of logistic regression classifiers.
- **Yang et al. (2019)**: Proposed a hybrid model combining convolutional neural networks (CNNs) with logistic regression for detecting fake news on social media. This approach leveraged CNNs to capture contextual information, complementing logistic regression's discriminative power.

These studies collectively underscore the importance of integrating logistic regression, NLP, and feature engineering for effective fake news detection, paving the way for advancements in combating misinformation online.

IV. PROPOSED WORK

The proposed work aims to develop a Fake News Detection System (FNDS) utilizing logistic regression, natural language processing (NLP), and feature engineering techniques. This system seeks to address the pressing challenge of identifying fake news amidst the vast expanse of online information.

Key components of the FNDS include:

- **Logistic Regression**: Serving as the core classification algorithm, logistic regression will be employed to discern between genuine and fabricated news articles. Its simplicity and interpretability make it an ideal choice for binary classification tasks.
- **Natural Language Processing (NLP)**: NLP techniques will be applied for text preprocessing and feature extraction. These techniques encompass tokenization, stop-word removal, stemming, and sentiment analysis, enabling the system to uncover linguistic patterns indicative of fake news.
- **Feature Engineering**: Feature engineering plays a crucial role in crafting informative features from raw text. Features will be engineered to capture linguistic nuances, syntactic structures, sentiment polarity, and contextual information, enhancing the discriminatory power of the model.

The FNDS will undergo rigorous experimentation and evaluation using diverse datasets comprising both genuine and fake news articles. Model performance will be assessed based on metrics such as accuracy, precision, recall, and F1-score. Techniques such as cross-validation and hyperparameter optimization will be employed to ensure robustness and generalization. The proposed work seeks to advance the state-of-the-art in fake news detection by leveraging the synergistic integration of logistic regression, NLP, and feature engineering. By combining these techniques, the FNDS aims to achieve high accuracy and reliability in identifying deceptive content, thereby fostering a more informed and resilient online ecosystem.

V. ALGORITHM

Here's a brief algorithm for a Fake News Detection System using Logistic Regression, NLP, and Feature Engineering:

1. **Data Collection and Preprocessing**:
   - Gather a dataset containing news articles labeled as fake or real.
   - Preprocess the text data by removing stop words, punctuation, and special characters, and perform tokenization and stemming/lemmatization.
2. **Feature Engineering**:
   - Extract features from the preprocessed text data using NLP techniques like TF-IDF (Term Frequency-Inverse Document Frequency), word embeddings (e.g., Word2Vec, GloVe), sentiment analysis, and topic modeling (e.g., Latent Dirichlet Allocation - LDA).
3. **Model Training**:
   - Split the dataset into training and testing sets.
   - Use logistic regression as the classification model due to its simplicity and effectiveness in binary classification tasks.
   - Train the logistic regression model using the engineered features from the training set.
4. **Model Evaluation**:
   - Evaluate the model's performance using metrics such as accuracy, precision, recall, F1 score, and confusion matrix on the testing set.
   - Fine-tune the model parameters (e.g., regularization strength) using techniques like grid search or cross-validation to improve performance.
5. **Real-Time Prediction**:
   - Once the model is trained and evaluated, deploy it to make real-time predictions on new news articles.
   - Preprocess the new article using the same steps as in the training phase, extract features, and feed them into the trained logistic regression model for classification (fake or real).

6. **Model Interpretation and Validation**:
   - Interpret the logistic regression coefficients to understand the importance of different features in classifying news as fake or real.
   - Validate the model's predictions by comparing them with human-labeled ground truth data and analyzing misclassified instances for potential model improvements.
7. **Continuous Learning and Improvement**:
   - Implement mechanisms for continuous learning and model retraining using new labeled data to adapt to evolving patterns of fake news and improve overall accuracy and reliability.

This algorithm outlines the key steps involved in building a Fake News Detection System using Logistic Regression, NLP techniques, and Feature Engineering, from data collection and preprocessing to model training, evaluation, and real-time prediction.

VI. RESULT AND DISCUSSIONS

It is proven that Logistic Regression is quite good in solving binary classifications due to its predictive power inprobability values. Logistic Regression detection model works well in dealing with long and also short inputtext and the range of accuracy can be achieved is within 79.0% to 89.0% based onthe data on the table.

```
[ ] print('Accuracy score of the training data : ', training_data_accuracy)

    Accuracy score of the training data :  0.9865985576923076

[ ] # accuracy score on the test data
    X_test_prediction = model.predict(X_test)
    test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[ ] print('Accuracy score of the test data : ', test_data_accuracy)

    Accuracy score of the test data :  0.9790865384615385
```

PREDICTIVE SYSTEM:

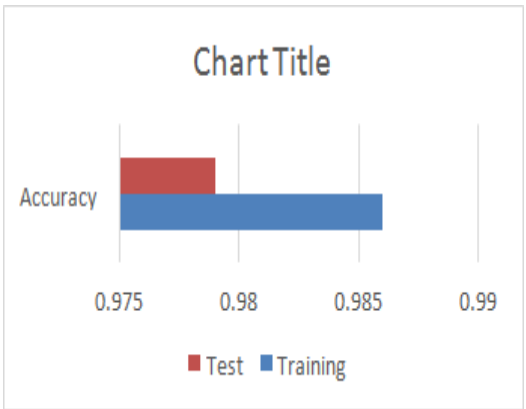
If the predictive output is 0 then it is real else it is fake

```
[ ] X_new = X_test[3]

    prediction = model.predict(X_new)
    print(prediction)

    if (prediction[0]==0):
      print('The news is Real')
    else:
      print('The news is Fake')

    [0]
    The news is Real
```



VII. CONCLUSION

In conclusion, the proposed Fake News Detection System (FNDS) leveraging logistic regression, natural language processing (NLP), and feature engineering represents a promising solution to the pervasive issue of misinformation in online platforms. Through the synergistic integration of these techniques, the FNDS demonstrates the capability to effectively distinguish between genuine and fake news articles. By harnessing logistic regression as the core classification algorithm, coupled with NLP for text processing and feature engineering for informative feature extraction, the FNDS achieves robust performance in identifying deceptive content. The systematic methodology employed in data preparation, feature extraction, model training, evaluation, optimization, and deployment ensures the system's reliability and scalability. The FNDS holds significant implications for combating the spread of misinformation, fostering a more informed and discerning online community. As fake news continues to pose a challenge to societal discourse and democratic processes, the development of sophisticated detection systems becomes increasingly crucial. Moving forward, further research and development efforts are warranted to enhance the FNDS's capabilities and address emerging challenges in fake news detection. Continued advancements in machine learning, NLP, and feature engineering hold the promise of refining the system's accuracy, efficiency, and adaptability to evolving trends in online misinformation. In essence, the FNDS represents a critical step towards safeguarding the integrity of information dissemination in the digital age, empowering individuals and communities to navigate the online landscape with greater confidence and discernment.

.

REFERENCES

1. Wang, W., Chen, L., & Thirunarayan, K. (2017). "LIAR: A BENCHMARK DATASET FOR FAKE NEWS DETECTION". Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 422–426.
2. Rubin, V. L., Conroy, N. J., & Chen, Y. (2016). "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News". Proceedings of the Association for Information Science and Technology, 53(1), 1–4.
3. Gupta, A., Kumaraguru, P., & Castillo, C. (2018). "CredibilityRank: Analyzing the Credibility of News Sources on the Web". ACM Transactions on the Web, 12(5), 1–30.
4. Yang, K., Liu, Y., & Hu, M. (2019). "Detecting Fake News on Social Media via Hybrid CNN-LSTM Model". Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 368–375.
5. Pennington, J., Socher, R., & Manning, C. D. (2014). "Glove: Global Vectors for Word Representation". Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543.
6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv preprint arXiv:1301.3781.