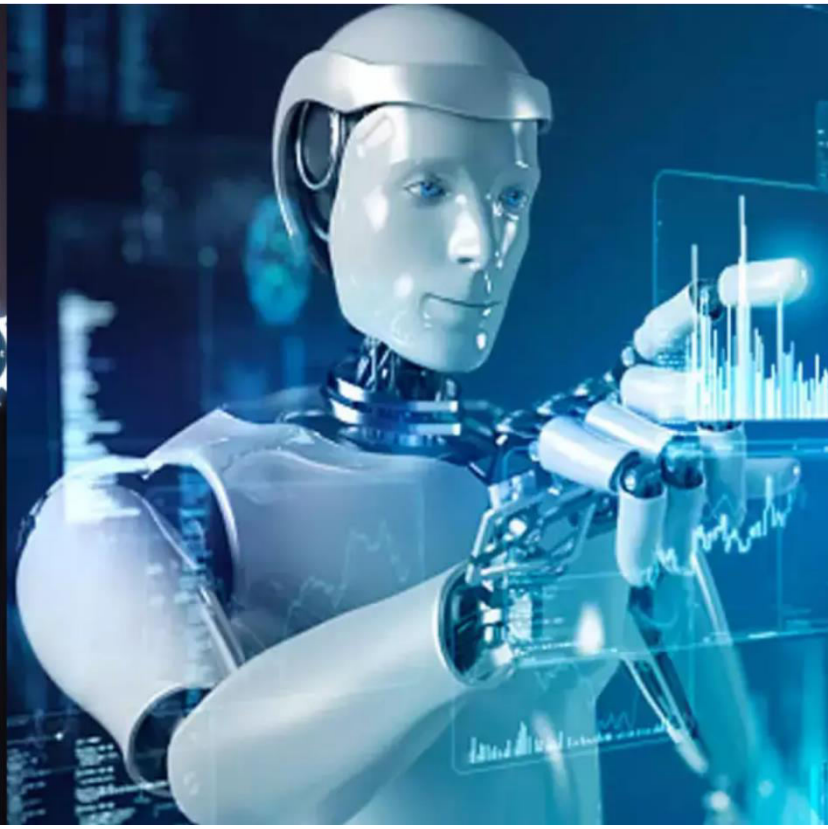
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*

**Impact Factor: 8.699**

**Volume 14, Issue 5, May 2025**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405197|

# Document Content Based Web Spam Detection Using KL-Divergence and Cosine Similarity Measure

**V.Nandhini, M.Chandrika**

P.G Student, Department of MCA, M.P.Nachimuthu M.Jaganathan Engineering College, Chennimalai, Erode, India

Assistant Professor, Department of MCA, M.P.Nachinuthu M.Jaganathan Engineering College, Chennimalai, Erode, India

**ABSTRACT:** Web spam is one of the main current problems of search engines because it strongly degrades the quality of the results. Many people become frustrated by constantly finding spam sites when they look for legitimate content. In addition, Web spam has an economic impact since a high ranking provides large free advertising and so an increase in the Web traffic volume. This project presents an effective spam detection system based on a classifier that combines new link-based features with language-model (LM)-based ones. These features are not only related to quantitative data extracted from the Web pages, but also to qualitative properties, mainly of the page links. In addition, it considers, for instance, the ability of a search engine to find, using information provided by the page for a given link, the page that the link actually at points. This can be regarded as indicative of the link reliability. Also, it checks the coherence between a page and another one pointed at by any of its links. Two pages linked by a hyperlink should be semantically related, by at least a weak contextual relation. Thus, it applies an LM approach to different sources of information from a Web page that belongs to the context of a link, in order to provide high-quality indicators of Web spam. In addition, cosine similarity between two web pages content is added to find coherence between a non-spam page and a new page to judge that page as spam or not.

## I. INTRODUCTION

The proliferation of online services has led to a significant increase in web-based scams, posing serious threats to individual users, organizations, and the broader digital economy. Scammers continuously evolve their techniques, making it increasingly challenging to distinguish between legitimate and fraudulent web activities using traditional detection methods. Consequently, there is a critical need for advanced, adaptive techniques capable of identifying subtle patterns and deviations characteristic of scam activities. This paper investigates a novel approach to web scam detection by utilizing Kullback-Leibler Divergence and Cosine Similarity as primary analytical tools. KL Divergence, an information-theoretic measure, quantifies the difference between two probability distributions and is particularly effective in detecting distributional shifts in feature spaces associated with scam behavior. By comparing feature distributions of known legitimate and potentially malicious web entities, KL Divergence provides a statistical basis for anomaly detection. Complementing this, Cosine Similarity measures the angular distance between high-dimensional feature vectors, offering a robust mechanism for assessing structural similarities between web activities. Cosine Similarity is particularly suited for high-dimensional, sparse data often encountered in web content analysis, making it an effective metric for identifying instances where fraudulent activities mimic legitimate patterns superficially but differ in their underlying structure. By integrating KL Divergence and Cosine Similarity, the proposed methodology aims to capture both the statistical divergence and the relational similarity between web entities, thus enhancing the precision and reliability of scam detection systems. The effectiveness of the proposed model is validated through experimental analysis on benchmark datasets, demonstrating its potential for practical deployment in dynamic web environments.

## II.LITERATURE SURVEY

The rapid increase in web-based scams has prompted extensive research into effective detection methodologies. Traditional scam detection techniques have primarily relied on rule-based systems, blacklist approaches, and supervised

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405197|

machine learning models. However, these methods often struggle with the dynamic and evolving nature of scams, necessitating more flexible and adaptive solutions. Several studies have explored statistical divergence measures for anomaly detection. Kullback Leibler (KL) Divergence has been widely used in fields such as network security, natural language processing, and fraud detection due to its ability to quantify differences between probability distributions. For example, Lee and Xiang (2001) utilized KL Divergence to detect network intrusions by identifying distributional changes in network traffic features. Their work demonstrated that significant divergence from normal patterns often indicates malicious activity, an approach that is similarly applicable to web scam detection. On the other hand, Cosine Similarity has been effectively used in text classification, document clustering, and web content analysis. It measures the angular similarity between high-dimensional vectors, making it robust to differences in magnitude and particularly suited for sparse data. Studies by Salton et al. (1975) in information retrieval highlighted Cosine Similarity's efficiency in measuring the closeness between textual entities. In the context of scam detection, Cosine Similarity has been employed to identify phishing emails and malicious URLs by comparing suspicious content against legitimate templates. More recent works have focused on combining multiple metrics to enhance detection accuracy. For instance, Gupta et al. (2019) proposed a hybrid model combining lexical feature analysis and behavioral similarity for phishing website detection, achieving improved performance over single-metric approaches. Their findings suggest that integrating statistical divergence with similarity measures can capture both distributional and structural characteristics of scams more effectively. Despite these advancements, limited research has directly combined KL Divergence and Cosine Similarity for web scam detection. Leveraging both measures offers a promising direction: KL Divergence captures how the underlying feature distributions shift, while Cosine Similarity assesses the relational closeness between samples. This dual approach has the potential to provide a more comprehensive detection framework, particularly in dynamic web environments where scammers continuously adapt their strategies. The proposed work builds upon these insights by developing a hybrid scam detection model that utilizes both KL Divergence and Cosine Similarity, aiming to address the limitations of existing single-method approaches and enhance detection robustness against evolving scam patterns

## III.METHODOLGY

The proposed methodology for web scam detection integrates statistical and similarity-based approaches to identify anomalies and deviation indicative of scam activities. The process involves four main stages: data collection and preprocessing, feature extraction, similarity and divergence computation, and classification.

**1. Data Collection and Preprocessing**

Data is collected from various web sources, comprising both legitimate and scam-related samples. Preprocessing steps include removing noise, normalizing data, handling missing values, and encoding categorical features. Textual information, such as web page content, URLs, and metadata, is tokenized and vector for further analysis.

**2. Feature Extraction**

Relevant features that distinguish between legitimate and scam web activities are extracted. These may include linguistic patterns (e.g., unusual word usage), structural features (e.g., URL length, domain age), behavioral features (e.g., user interaction patterns), and content-based metrics (e.g., frequency of suspicious terms). Feature vectors are generated for each web entity.

**3. Similarity and Divergence Computation**

Two complementary measures are applied to the extracted feature vectors:

KL Divergence Calculation:

The probability distributions of features from legitimate and suspected scam samples are estimated. KL Divergence is then computed to quantify the difference between the two distributions. A higher divergence value suggests greater deviation from legitimate behavior, indicating potential scam activity.

The KL Divergence is computed as:

$$D_{KL}(P \parallel Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)}$$

Cosine Similarity Computation:

Cosine Similarity is calculated between feature vectors to measure the closeness of suspicious samples to legitimate web behavior. A lower similarity score indicates a greater likelihood of scam behavior.

Cosine Similarity is computed as:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405197|

\text{cos}(\theta) = \frac{A \C. B}{\|A\| \|B\|}

**4. Classification and Decision Making**

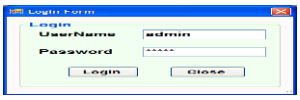
Based on the KL Divergence and Cosine Similarity scores, a composite risk score is generated for each web entity. Threshold-based or machine learning classification techniques (such as SVM, Random Forest, or Logistic Regression) are used to label the web entities as "legitimate" or "scam." Entities with high KL Divergence and low Cosine Similarity relative to legitimate samples are classified as scams.

**5. Evaluation Metrics**

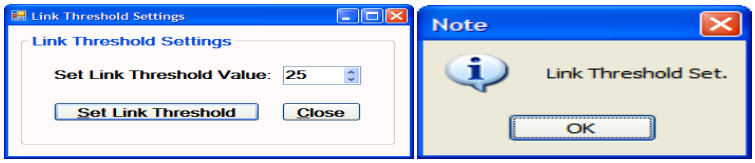
The performance of the proposed detection system is evaluated using metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), ensuring a comprehensive assessment of detection effectiveness.

## IV. RESULTS

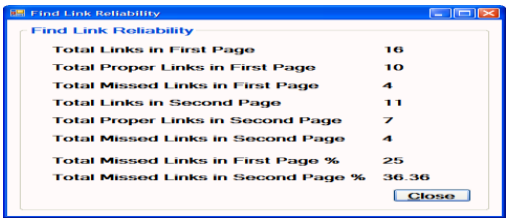
### LOGIN FORM



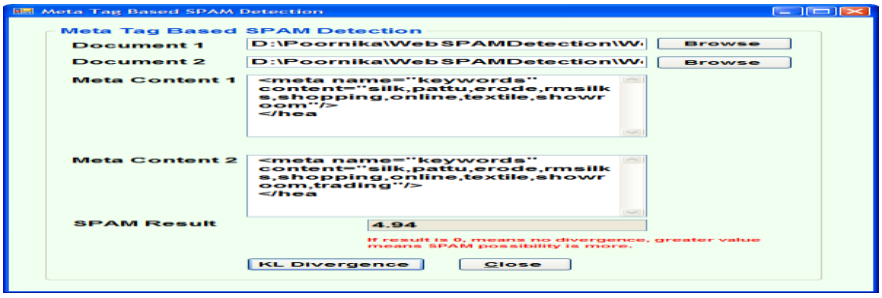

### LINK THRESHOLD SETTING




### FIND LINK RELIABILITY




### META TAG



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

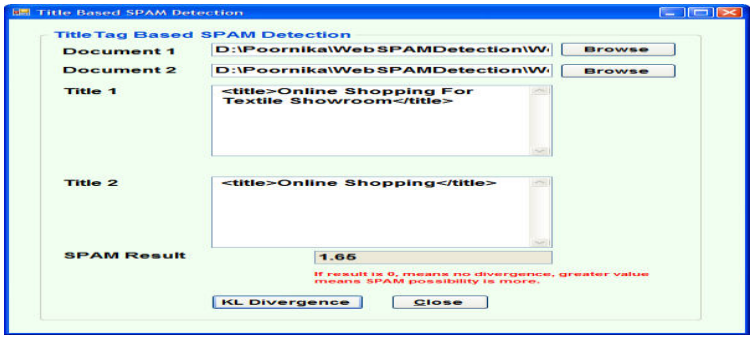
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

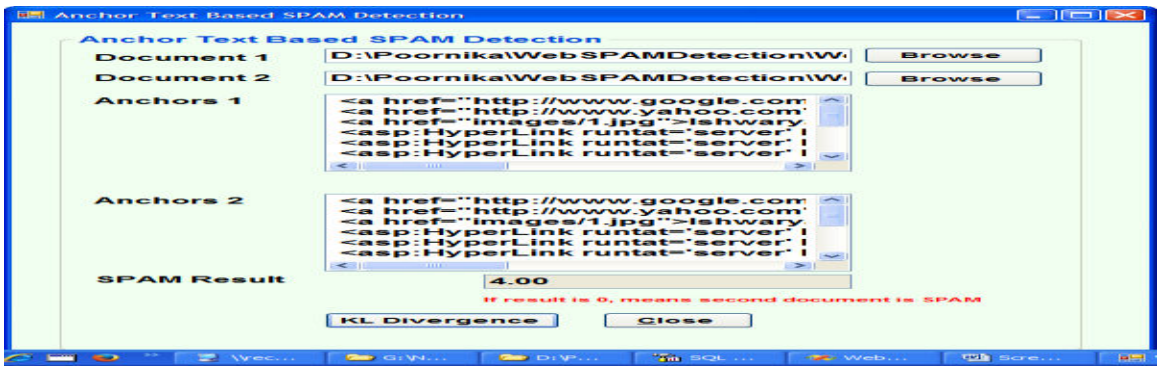
Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405197|

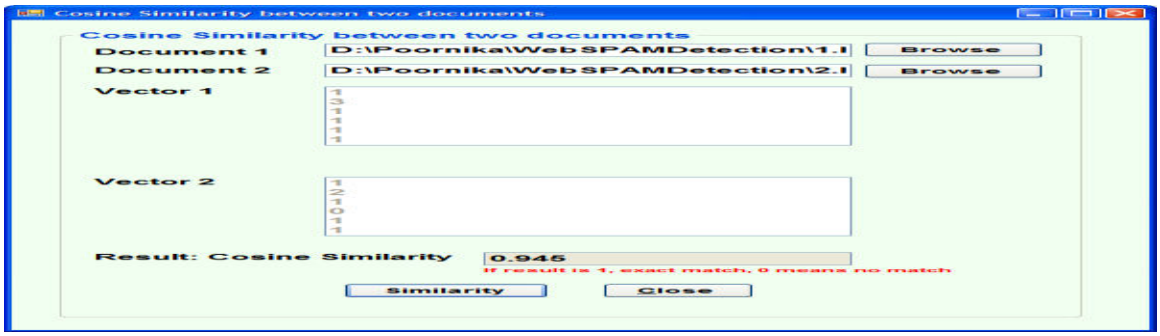
TITLE TAG




ANCHOR/TEXT BASED




COSINE SIMILARITY BETWEEN TWO DOCUMENTS




## V. CONCLUSION

In this project, the proposed a new methodology is used to detect spam in the Web, based on an analysis of QLs and a study of the divergence between linked pages. To use QL's and the LM features effectively, we proposed a robust classifier based on a cost-sensitive algorithm. It has evaluated the presented methodology using the public datasets and it focus on the F-mea-sure, using the proposed features in a separate and also in a combined way. It has been proven that QL features have obtained better results than pre-computed content and link-based features, even with many fewer features. In addition, when we combine the four sets of features and we apply them to WEB-SPAM-UK2006 and WEBSPAM-UK2007 datasets, the system detects sum amount of the spam domains, with an F-measure process, respectively. Thus, an improvement of the measure of in certain range level respectively, is obtained. Therefore, the comparisons with pre-computed features show that the proposed methodology yields much better performance, indicating that LMs and QLs can be used to detect Web spam effectively.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405197|

## REFERENCES

### BOOK REFERENCES

1. Alistair Mc Monnies, "**Object-Oriented Programming in Visual C#".** Person Education and ISBN: 81-297-0649-0, First Indian Reprint 2004.
2. Jittery R Shapiro, "**The Complete Reference Visual C# .NET**" Edition 2002, Tata McGraw-Hill, Publishing Company Limited, New Delhi.
3. Roger S. Pressman "**Software Engineering**" Tata McGraw-Hill, Publishing Company Limited (1987).
4. Robert D Schneider, Jetty R Garbus, "**Optimizing SQL Server**", Second Edition, Pearson Education Asia, ISBN: 981-4035-20-3 (2000).
5. Francesco Balena, "**Programming Microsoft Visual C# 2005: The Language",** Microsoft Press.
6. David I Schneider, **"Introduction to Programming with Visual C# .Net 2005",** Prentice Hall.

### WEB REFERENCES

1. http://www.codeproject.com
2. http://c-sharpcorner.com
3. http://msdn.microsoft.com

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details