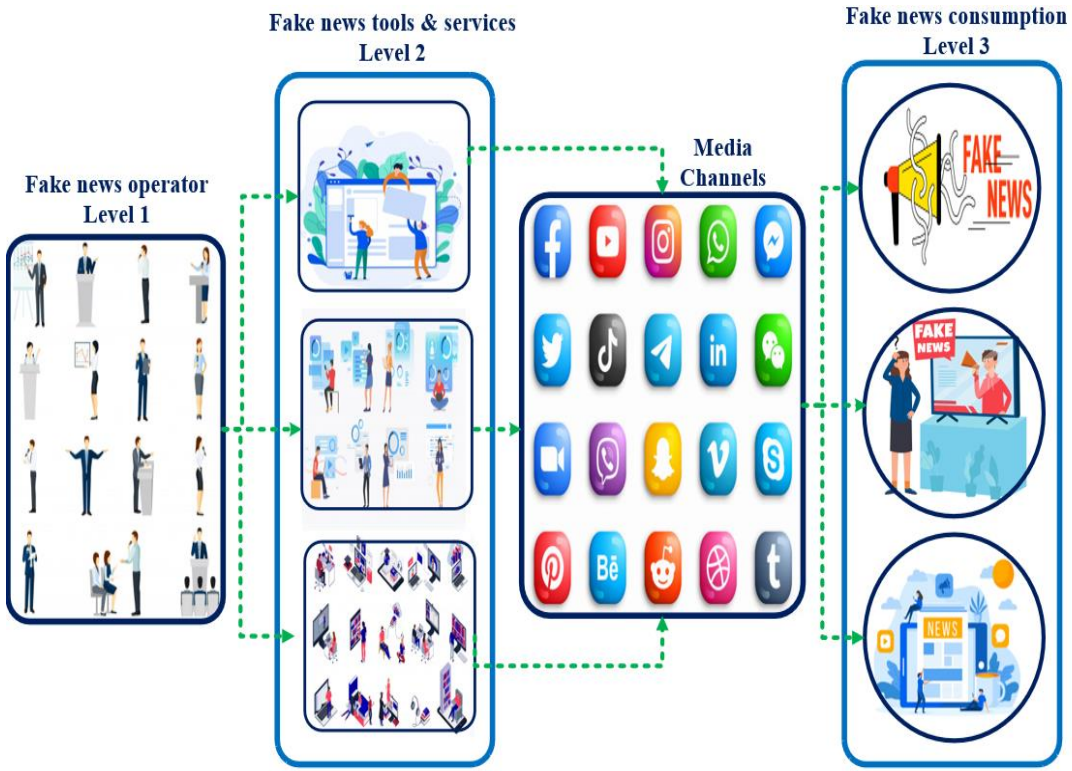
Fake Review Classification using Supervised Machine Learning

Abstract

- The revolution of social media has propelled the online community to take advantage of online reviews for not only posting feedback about the products, services, and other issues but also assists individuals to analyze user's feedback for making purchase decisions, and companies for improving the quality of manufactured goods.
- However, the propagation of fake reviews has become an alarming issue, as it deceives online users while purchasing and promotes or demotes the reputation of competing brands.
- In this work, we propose a supervised learning-based technique for the detection of fake reviews from the online textual content.
- The study employs machine learning classifiers for bifurcating fake and genuine reviews. Experimental results are evaluated against different evaluation measures and the performance of the proposed system is compared with baseline works.

Fake News

- There has been a rapid increase in the spread of fake news in the last decade, most prominently observed in the 2016 US elections .
- Such proliferation of sharing articles online that do not conform to facts has led to many problems not just limited to politics but covering various other domains such as sports, health, and also science.
- One such area affected by fake news is the financial markets, where a rumor can have disastrous consequences and may bring the market to a halt.



Fake News Detection

- It is an important task to design a system that can detect and classify text into fake (spam) and real (non-spam) reviews.
- Fake news detection at an earlier stage will assist the online community in making decisions correctly and analyzing customer feedback efficiently.
- Prior works on the fake review detection have used supervised machine learning (ML) and lexicon-based techniques.

Research Questions

- RQ1. How to implement a supervised ML technique for classifying text into spam and non-spam?
- RQ2. How to evaluate the efficiency of the supervised ML system for classifying text into non-spam (genuine) and spam (fake) classes?
- RQ3. What is the efficiency of the proposed technique concerning similar studies for classifying spam reviews efficiently?

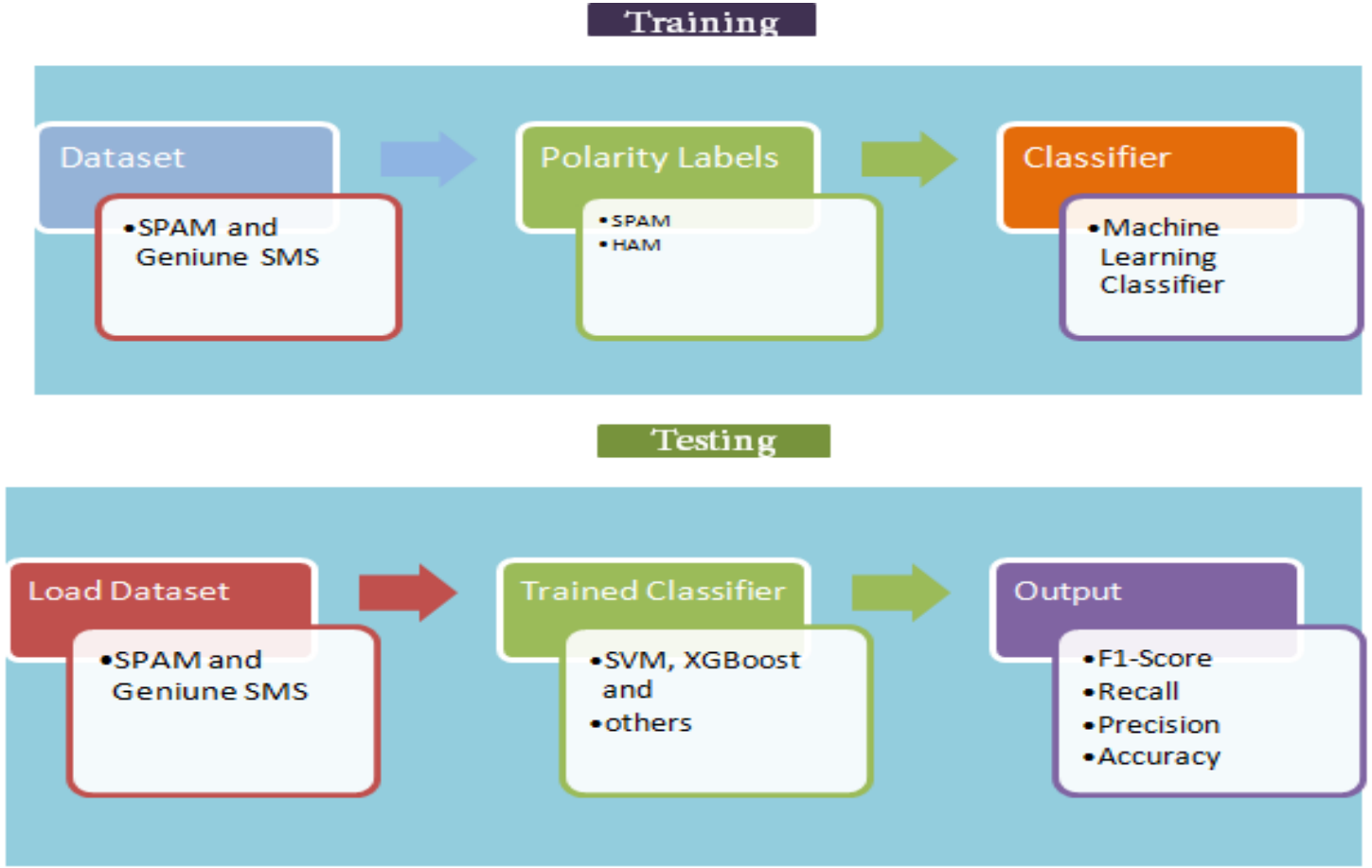
Objectives

- To implement supervised ML technique for classifying text into spam and non-spam.
- To evaluate the efficiency of the supervised ML method for the classification text into spam and non-spam classes.
- To evaluate the efficiency of the proposed technique concerning similar studies for classifying spam reviews efficiently.

Research Contributions

- Classification of text into binary classes (spam and non-spam) by applying supervised ML technique.
- Evaluating the efficiency of the proposed system for efficiently classifying text into spam and non-spam classes.
- Comparison of the proposed system with the state-of-the-art techniques developed for spam detection.

Architecture of Proposed System for Spam Detection



Proposed System

- The supervised learning technique for classifying spam and genuine reviews starts by entering input review, pre-processing the review, and finally classifying it as fake (SPAM) and genuine (HAM) using the SVM classifier.
- The dataset is split into a train (80%) and test (20%) modules. At the training stage, labeled data is provided to the classifier.
- Once the training stage of classifiers is finished, the classification of ML classifiers is validated through the assessment of the rest of the testing data.
- The obtained result is evaluated through multiple measures like Precision, F1-measure, Accuracy, and Recall.

Results and Discussion

Experimental results of different ML classifiers.

Classifier	F1-score %	Recall %	Precision %	Accuracy %
K-Nearest Neighbour	0.89	0.89	0.89	83.25
Support Vector Machine	0.99	0.99	0.99	98.92
Decision Tree	0.97	0.97	0.97	97.31
XGBoost	0.97	0.97	0.97	96.77
Logistics Regression	0.97	0.97	0.97	97.13
Navies Bayes	0.98	0.98	0.98	97.67
Random Forest	0.97	0.97	0.97	96.5

Results and Discussion

Cross validation of different classifiers

Classifiers	Mean Accuracy	Standard Deviation	Mean Precision macro	Standard Deviation	Mean Recall Macro	Standard Deviation	Mean F-I Macro	Standard Deviation
Random Forest	0.971	0.005	0.984	0.005	0.895	0.028	0.935	0.012
SVM	0.988	0.004	0.988	0.006	0.959	0.016	0.988	0.004
KNN	0.915	0.005	0.955	0.002	0.682	0.018	0.743	0.021
Logistic Regression	0.971	0.005	0.979	0.007	0.896	0.017	0.931	0.013
XB Boost	0.968	0.006	0.974	0.008	0.891	0.024	0.930	0.019
Decision Tree	0.965	0.005	0.942	0.006	0.909	0.019	0.926	0.011
NB	0.968	0.006	0.938	0.013	0.911	0.025	0.925	0.015

Results and Discussion

Comparison with Baselines Result

Studies	Technique	Results
Pragna & RamBai	SVM	96.23% (Accuray)
Renuka *et al.*	MLP	93 % (Accuray)
Our Work	Machine Learning Classifier: SVM	98.92% (Accuracy) 99% (Recall) 99% (Precision) 99% (F-Measure)

Future Directions

- A balanced dataset can improve the classification result of the ML classifiers.
- Different techniques for segmentation of dataset e.g. hold-out cross-validation, cross-validation, stratified sampling, and a few more, can also be applied to evaluate the classification performance of ML classifiers.
- Other feature engineering techniques, such as word embedding can also be investigated for better results.
- Increasing the size of the dataset can also produce more promising results.
- Blockchain technology can be used to identify the source of the fake news.