# Paper1

## TITLE:

[Content Based Spam Detection In Short Text Messages With Emphasis On Dealing With Imbalanced Datasets](https://ieeexplore.ieee.org/abstract/document/8697372)

## INTRO:

The research on spam detection has explored various machine learning techniques to filter spam messages from text. Different classifiers, including **SVM, Naïve Bayes, Random Forest, kNN, and Artificial Neural Networks**, have been tested, with **Random Forest** achieving high accuracy (97.88% for SMS spam and 93.43% for Twitter spam). While most research has relied on the **Bag of Words model**, efforts to address dataset imbalance and identify key discriminating features remain limited.

A diagram of a model

Description automatically generatedDATASET:

The dataset is **imbalanced**.

## PROPOSED WORK:

### 1-Pre-Processing Module:

**a-Tokenization**: Splits text into smaller units called tokens, which correspond to words.

**b-Stop Word Removal**: Filters out common words (e.g., "in," "if," "for") that do not contribute to classification.

c-**Stemming**: Reduces words to their base form by chopping off endings to unify different word variations.

2-feature extraction:  
a-**N-grams**: Continuous sequences of *n* words in training data (unigram, bigram, trigram).

* Unigrams act as strong discriminators.
* Bigrams and trigrams help capture contextual information.

b-**Numeric String**: Identifies sequences of numbers and analyzes their probability of indicating spam.

c-**Message Length**: Measures the number of characters in a message to assess the likelihood of it being spam based on length.

*Multiple papers also used TF-IDF so consider using it.*

### 3- Dataset Balancing Module:

To deal with imbalanced data, the proposed approach experiments with **SMOTE**

### 4- Model Generation using Machine Learning Algorithm

The model generation process involves filtering n-grams based on occurrence thresholds to handle sparsity. Four machine learning classifiers**: Adaboost, J48, SVM, and Bagging**—are used for experimentation. Since no single solution effectively handles imbalanced datasets, multiple classifiers are tested, and parameters are fine-tuned using the SMOTE approach. An iterative process helps identify the best-performing classifier for spam detection.

### 5-results:

Comparatively SVM exhibits a better performance across the dataset.

A table with numbers and text

Description automatically generatedA graph of different colored bars

Description automatically generated with medium confidence

# Paper2

## TITLE:

[An Integrative Data-Driven Architecture for Online Social Network Spam Detection Using Data Balancing and Machine Learning Methods](https://www.researchgate.net/profile/Ijeasm-Journal/publication/388174907_An_Integrative_Data-Driven_Architecture_for_Online_Social_Network_Spam_Detection_Using_Data_Balancing_and_Machine_Learning_Methods/links/678d0d9f95e02f182e9fb98b/An-Integrative-Data-Driven-Architecture-for-Online-Social-Network-Spam-Detection-Using-Data-Balancing-and-Machine-Learning-Methods.pdf)

## INTRO:

Using the **SMOTEENN** sampling approach, which combines SMOTE and Edited Nearest Neighbors (ENN) to produce balanced datasets. Advanced deep learning classification techniques are then applied to these datasets in order to differentiate between spam and authentic content (ham).

NOTE:  
this paper work with tabular data not content so we will just take the SMOTEENN idea from it