Spam Classifier

Shreya Hunur, Rohit Yadav, Hashmitha Katta

San Jose State University

May 2022

***Abstract*— Because of its global accessibility, relatively quick message transfer, and low sending cost, SMS is one of the most popular and widely used modes of communication. With technological advancements and an increase in content-based advertising, the use of Short Message Service (SMS) on phones has increased to the point where devices are occasionally flooded with spam SMS. Spam is a term that is commonly used to describe junk or unsolicited messages or information. As a result, a spam SMS is any junk message delivered to a mobile device via text messaging. The danger is the same whether the spam is sent via email or SMS. Spam may result in the disclosure of personal information, invasion of privacy, or unauthorized access to mobile device data.**

Most Internet users openly despise spam, but enough of them respond to commercial offers that spam remains a viable source of income for spammers. While most users want to do the right thing to avoid and eliminate spam, they require clear and simple guidelines on how to behave. Despite all the measures taken to eliminate spam, it has not been completely eradicated. Furthermore, if the countermeasures are overly sensitive, even legitimate SMS will be deleted.

Keywords—Short Message Service, Machine Learning, UCI, Random Forest

1.Introduction

Users now have personal and confidential information such as contact lists, credit card numbers, photographs, passwords, and much more stored in their smartphones, making them vulnerable to cyber-attacks via spam SMS. This allows hackers engaged in unethical activities to access smartphone data without the end-knowledge, jeopardizing the user's privacy. This could result in both monetary and functional losses. Spam messages appear to be on the rise, causing not only annoyance but also critical data loss for some users. Aside from that, SMS spam can be a driving force for malware and keyloggers.

Spammers use these spam messages to promote their utilities or businesses. Users may sometimes suffer financial losses as a result of spam messages. The widespread distribution of this type of problem, as well as less effective security control measures, has inspired many researchers to develop a set of techniques to assist simple formulae, giving them an advantage over other techniques in particular.

Machine Learning is a technique that allows machines to learn from past data and estimate future data. Machine learning and deep learning can now be used to tackle most real-world issues in a variety of fields, including health, security, market research, and more.We intend to compare various classifying techniques on various datasets gathered from previous research works and evaluate them based on their accuracies, precision, recall, and ROC Curve. A comparison of traditional machine learning techniques and deep learning methods will be carried out.

2. Research Methodology

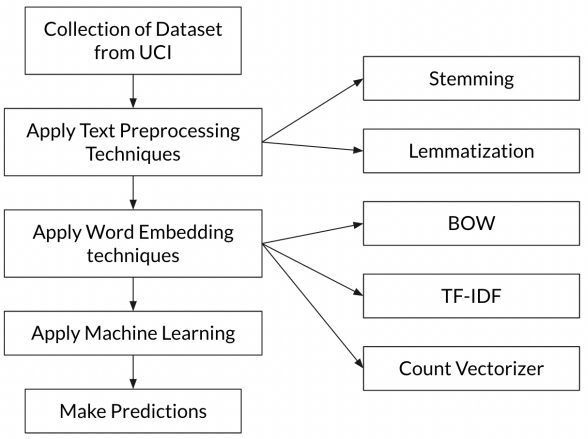

Fig. 1. Proposed approach for SMS spam detection.

2.1 Dataset

The UCI Machine Learning Repository has datasets for machine learning techniques. The SMS Spam Collection (hereafter the corpus) is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according to ham (legitimate) or spam.

The files contain one message per line. Each line is composed of two columns: one with a label (ham or spam) and other with the raw text.

Number of Training Samples	Number of Testing Samples	Total
4180	1394	5574

Fig. 2. Details of the Dataset.

2.2 Pre-Processing

2.1.1 Data cleaning:

Data cleaning is a very crucial process in NLP. In this only alphabetic characters are extracted from text removing the punctuation and numbers, and then all characters are converted into lowercase. This cleaned text will be then used in further processing.

2.1.2 Tokenization :

Tokenization is breaking the raw text into small chunks.Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding

the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.

2.1.3 Removing stop words:

Stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc. Stop words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information.

2.1.4 Stemming and Lemmatization:

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. However, the two words differ in their flavor.

Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

In Fig.3. we can see how the text looks after stemming and lemmatization.

clean	stemmed	lemmed
[go, jurong, point, crazy, available, bugis, n...	[go, jurong, point, crazi, avail, bugi, n, gre...	[go, jurong, point, crazy, available, bugis, n...
[ok, lar, joking, wif, u, oni]	[ok, lar, joke, wif, u, oni]	[ok, lar, joking, wif, u, oni]
[free, entry, 2, wkly, comp, win, fa, cup, fin...	[free, entri, 2, wkli, comp, win, fa, cup, fin...	[free, entry, 2, wkly, comp, win, fa, cup, fin...
[u, dun, say, early, hor, u, c, already, say]	[u, dun, say, earli, hor, u, c, alreadi, say]	[u, dun, say, early, hor, u, c, already, say]
[nah, dont, think, goes, usf, lives, around, t...	[nah, dont, think, goe, usf, live, around, tho...	[nah, dont, think, go, usf, life, around, though]

Fig. 3. Input text after stemming and lemmatization.

2.1.5 Bag of Words

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things: A vocabulary of known words. A measure of the presence of known words.

2.1.5 TF-IDF

TF-IDF stands for “Term Frequency — Inverse Document Frequency”. This is a technique to quantify words in a set of documents. We generally compute a score for each word to signify its importance in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.

TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)

Terminology:
t: term (word)
d: document (set of words)
N: count of corpus
corpus: the total document set

2.1.6 Outlier Detection and method used to treat Outliers:

Outliers are extreme values on data that differ from other observations.. In other terms, an outlier is a data point that deviates from the sample's overall pattern.

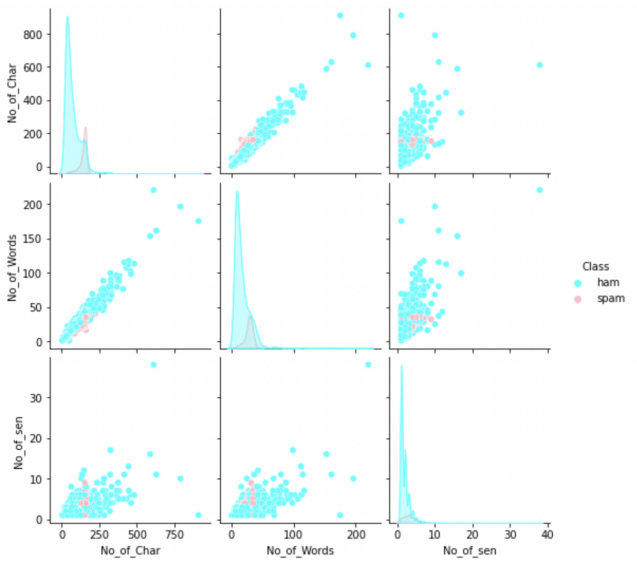

Fig. 4. Outlier Detection.

Observation : From the pair plot, we can see a few outliers all in the class ham. This is interesting as we could put a cap over one of these. As they essentially indicate the same thing i.e. the length of SMS.

Since the data we are working with is input text, the outliers were dropped.

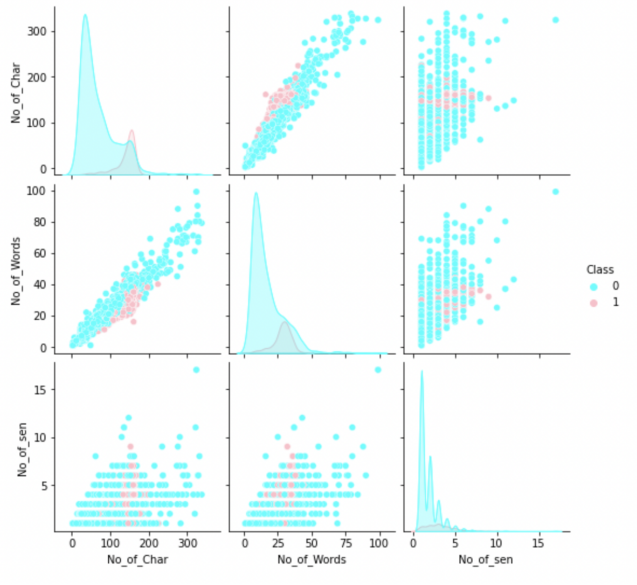

Fig. 5. PairPlot after removing Outliers.

Observation : Here, we can observe that the graphs are little zoomed in as we dropped a few outliers.

2.3 Data Visualization

Data visualization is the graphical representation of information and data. Using visual features like charts, graphs, and maps, data visualization tools make it simple to explore and comprehend trends, outliers, and patterns in data.

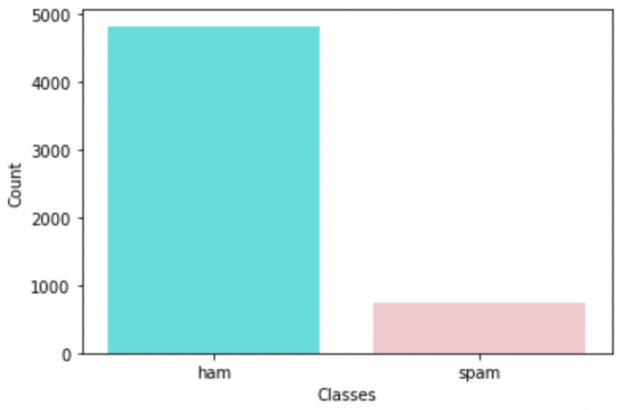

Fig. 6. Spam Distribution.

Observation : Number of legitimate messages(ham) are much more than spam messages.

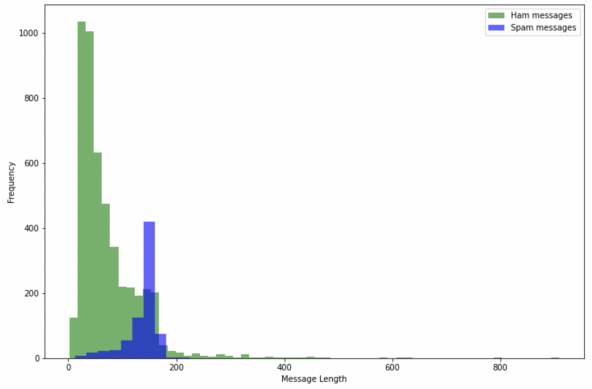

Fig. 7. Message length vs Frequency.

Observation: We can observe that spam messages are generally longer than ham messages: Bulk of ham has length below 100, for spam it is above 100.

A deep learning model for SMS spam detection was proposed in this work. For our experiments, we used a UCI dataset.

For visualizing the data, we did feature engineering. The length of each message was calculated.

The spam classification in the dataset consists of output in the form of text: ‘spam’ and ‘ham’.

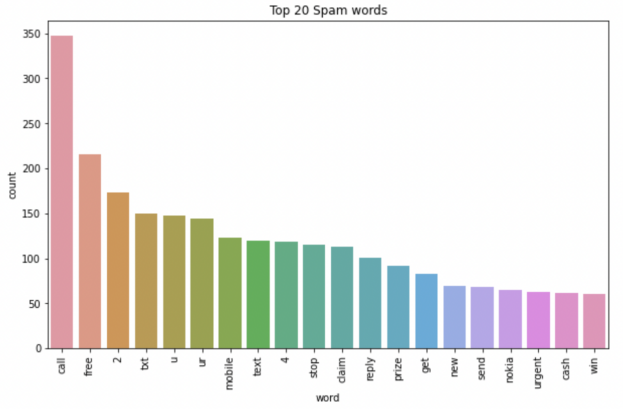

Fig. 8. Top 20 spam words after preprocessing.

Observation : We observe that words like call, free and misspelled words or abbreviations are common in spam messages.

3.METHODS

While training our models, we performed hyperparameter tuning using GridSearchCV.It is a technique for finding the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique where the model as well as the parameters are inputs.. After extracting the best parameter values, predictions are made.

3.1 Multinomial Naive Bayes

Naive Bayes is a probabilistic algorithm based on the Bayes Theorem that is used in data analytics for email spam filtering**.** The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.

In our case, we have a frequency as a feature. In such a scenario, we use multinomial Naive Bayes. It ignores the non-occurrence of the features. So, if you have frequency 0 then the probability of occurrence of that feature will be 0 hence multinomial naive Bayes ignores that feature.

3.2 Random Forest

Random forest is a meta-learner composed of many individual trees**.** Each tree votes on the overall classification of the data set, and the random forest method chooses the classification with the most votes. Replacement sampling is used to build each decision tree from a random subset of the training dataset.That is, some entities will appear more than once in the sample, while others will not. To determine how to divide the dataset optimally at each node, each decision tree is built using a model based on a separate random subset of the training dataset and a random subset of the available variables. Each decision tree is built to its maximum size without any pruning. When the Random forest decision tree models are combined, they produce the final ensemble model, in which each decision tree votes for the outcome and the majority wins.

3.3 SVC

The Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression tasks. It is, however, mostly used in classification problems. Each data item is plotted as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a specific coordinate in the SVM algorithm. Then, we perform classification by locating the hyperplane that best distinguishes the two classes**.**

3.4 K-Neighbours

The KNN algorithm is a simple and effective classification algorithm. The main issue with this algorithm is determining what the k value will be at the start. The algorithm's performance varies depending on the k value chosen.KNN does not have a separate training stage followed by a stage in which the labels for the test data are predicted based on the trained model. Rather, the features of each test data item are compared in real time with the features of each training data item, and the K nearest training data items are chosen, and the most frequent class among them is assigned to the test data item.

4.Comparison

4.1 Performance Metrics

4.1.1 Accuracy :

Accuracy is not a reliable metric for a classifier's true performance when the number of samples in different classes varies greatly (unbalanced target), as it will produce misleading results.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total of all cases to be predicted}}$$

4.1.2 Precision :

A classification model's ability to identify only relevant data points. Precision is defined mathematically as the number of true positives divided by the number of true positives plus the number of false positives.

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

4.1.3 Recall :

A model's ability to find all relevant cases within a data set. We define recall mathematically as the number of true positives divided by the total number of true positives plus the number of false negatives.

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

4.1.4 F1 score :

The F1 score is the harmonic mean of precision and recall in the following equation, which takes both metrics into account.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

We use the harmonic mean instead of a simple average because it punishes extreme values. A classifier with a precision of 1.0 and a recall of 0.0 has a simple average of 0.5 but an F1 score of 0.

4.1.5 AUC-ROC :

The AUC - ROC curve is a performance metric for classification problems at various threshold levels. AUC represents the degree or measure of separability, while ROC is a probability curve. It indicates how well the model can distinguish between classes.

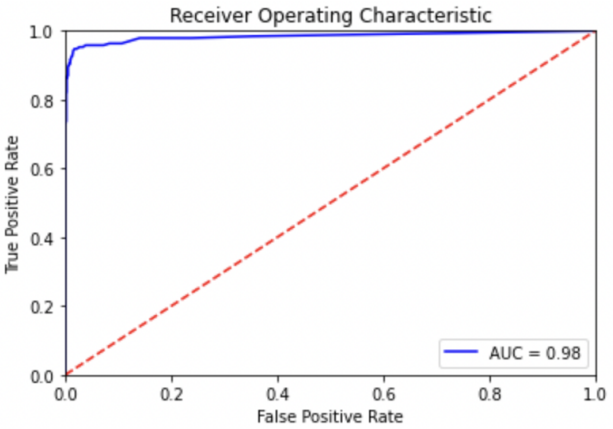

Fig. 9. ROC curve.

4.2 Comparison

Among all the models we implemented, the Random Forest Classifier performed exceptionally well. As you can see from Fig.10. that the F1 score for Random Forest is very high.

	Accuracy	Precision	Recall	F1 score	ROC AUC
Random Forest	0.97	1.0	0.97	0.99	0.99
SVC	0.97	1.0	0.97	0.98	1.0
KNN	0.89	1.0	0.90	0.94	0.95
Naive Bayes	0.96	1.0	0.96	0.98	0.99

Fig. 10. Comparison of performance metrics of different models.

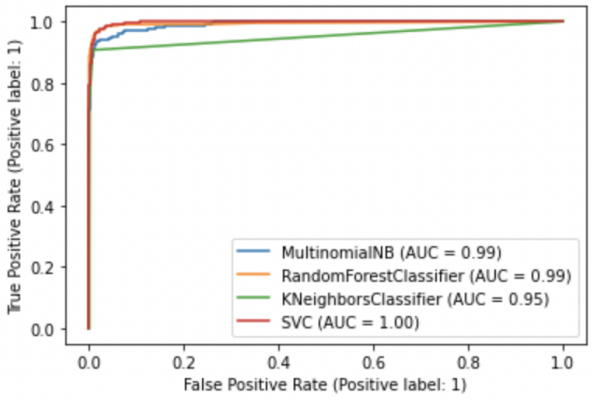

Fig. 11. Comparison of ROC curves of different models.

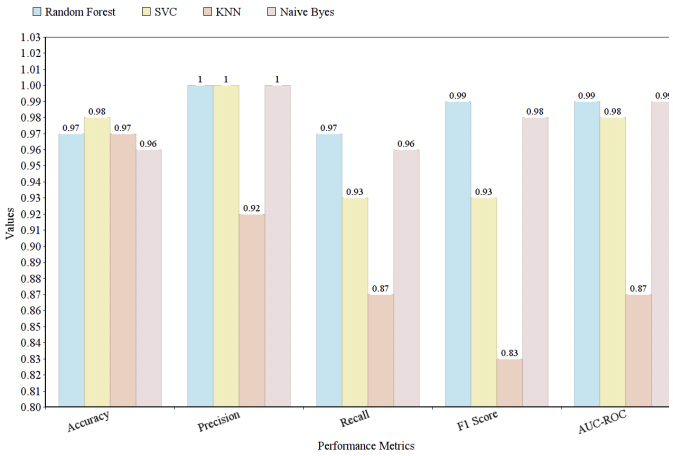

Fig. 12. Comparison of performance metrics of different models.

5.Conclusion

This article discusses spam filtering techniques based on machine learning algorithms. For our experiments, we used a UCI dataset. Only this dataset was used to test our model. Bag of Words, Count vectorizer and TF-IDF are three distinct word embedding approaches that we used. In terms of accuracy score, TF-IDF with Random Forest classification method surpasses other algorithms in the trial. However, because the dataset is unbalanced, it is not enough to judge the performance just on accuracy; the precision, recall, ROC score and f1-score of the methods must also be considered. After further testing, the RF algorithm still manages to deliver good precision and f1-score, with precision of 0.97, ROC score of 0.99 and f1-score of 0.99 . Based on the characteristics employed, different algorithms will provide different performances and results. Adding extra characteristics may aid classifiers in improving training data and performance. We'll test our model on a variety of datasets in the future.

References

[1] D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques," 2011 International Conference on Process Automation, Control and Computing, 2011, pp. 1-7.

[2] Shrawan Kumar Trivedi, "A Study of Machine Learning Classifiers for Spam Detection", 2016 4th International Symposium on Computational and Business Intelligence,2016.

[3] Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Deepika Koundal, Bader Alouffi, and Tariq Shah "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges", 3 February 2022 Hindawi Security and Communication Networks Volume, 2022.

[4] Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal & Pulkit Mehndiratta , "A Comparative Study of Spam SMS Detection using Machine Learning Classifiers", Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018.

[5] https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1