SPAM MESSAGES FILTER

(USINNG SVM AND NAÏVE BAYES):TUSHAR SINGH

SMS filters importance

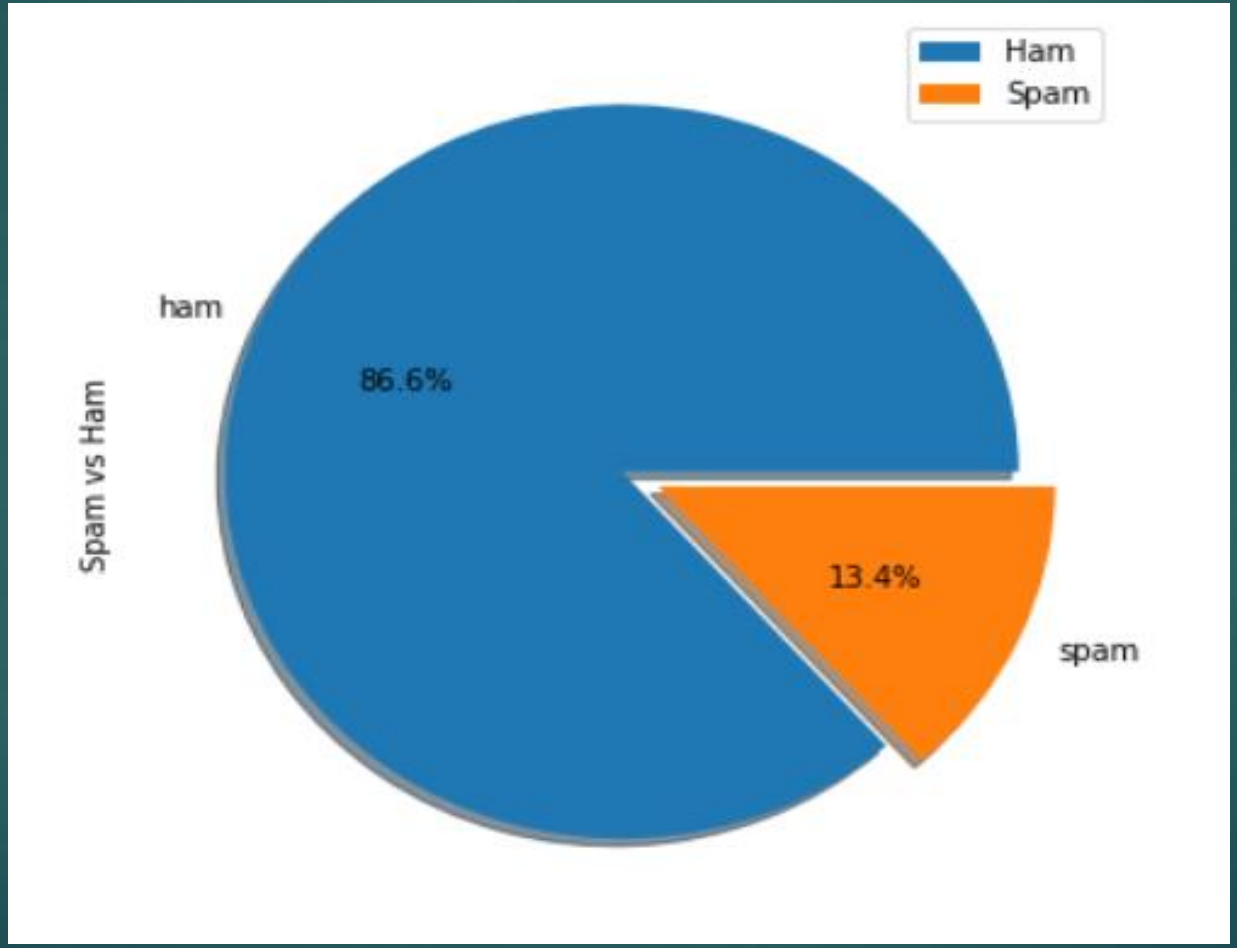
- Due to the spread of fake messages and spam messages by spammer

  theres a lot of clickbait so filtering it using ML would just make our life less buggy
- Due to this there are many fake messages which lets to the money scams through messages
- This not only saves our data and time but make our digital life more secure

Data Acquisition and Cleaning

- The data has been acquired by Kaggle dataset UCI_MACHINE LEARNING DATASET . The SMS Spam Collection 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 acording being ham (legitimate) or spam.
- The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.
- Data Cleaning (Removing unimportant data/ Stopwords/ Stemming)
- Converting data into a model usable format (Bag of words/ Tfidf Model)

Proportion of Ham(Legit) and Spam in the Class



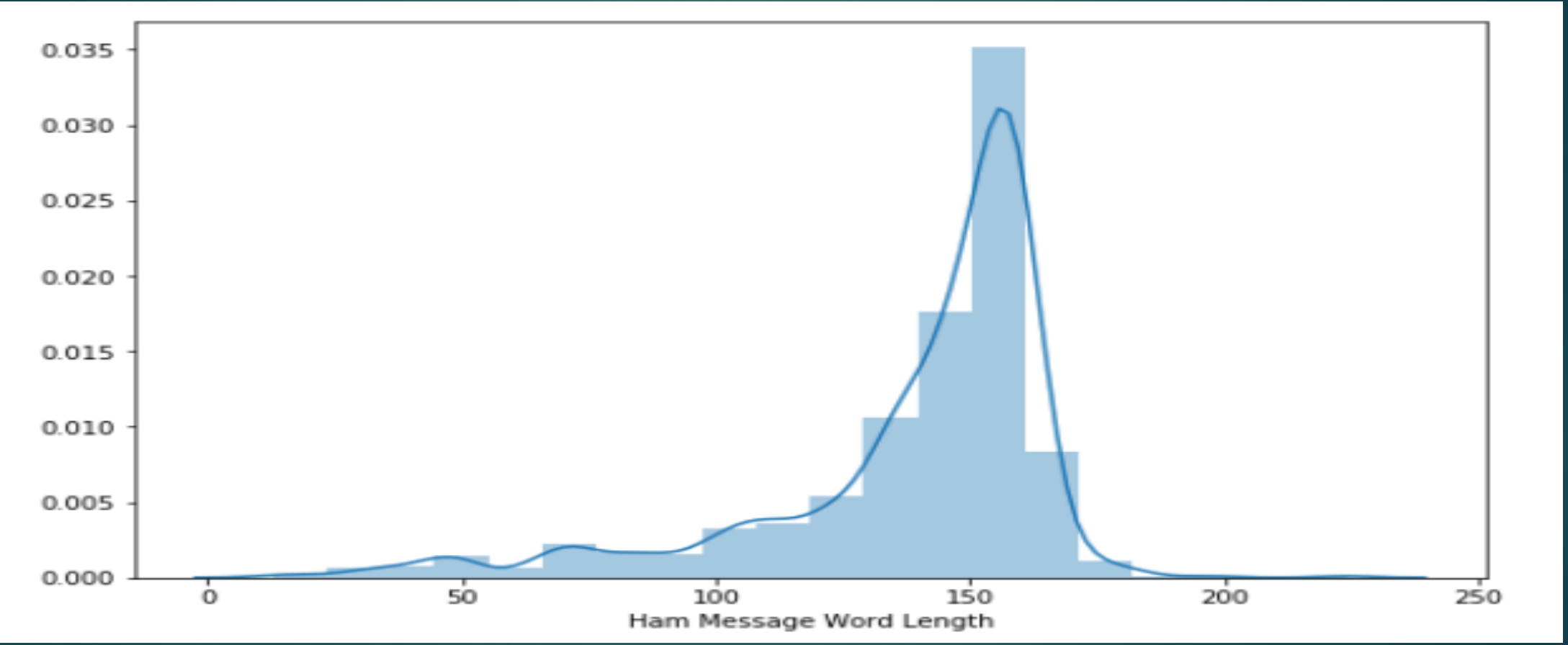
Features selection for training

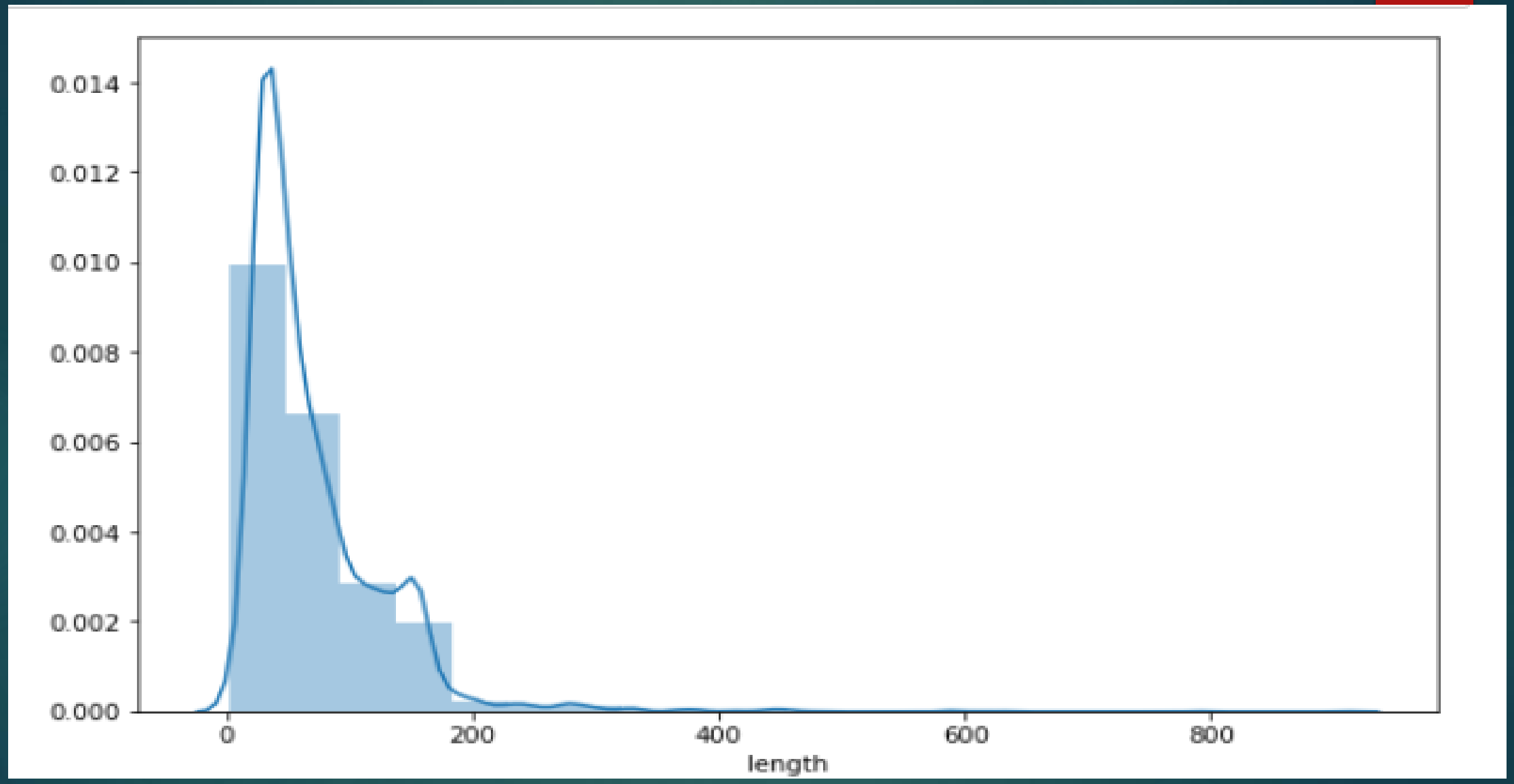
- In the data set we have renamed the v1 and v2 to class and text columns
- The length of the messages ha been

text
Sorry, I'll call later
I cant pick the phone right now. Pls send a message
Ok...
Your opinion about me? 1. Over 2. Jada 3. Kusruthi 4. Lovable 5. Silent 6. Spl character 7. Not matured 8. Stylish 9. Simple Pls reply..
ur lovable bcums angry wid u, dnt take it seriously.. Coz being angry is d most childish n true way of showing deep affection, care n luv!.. kettoda manda... Have nice day da.
ll our customer service representative on FREEPHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed å£1000 cash or å£5000 prize!
Okie
wly.? GOD,I LOVE YOU & I NEED YOU,CLEAN MY HEART WITH YOUR BLOOD.Send this to Ten special people & u c miracle tomorrow, do it,pls,pls do it...
n My WORLD 7th You 6th Ur style 5th Ur smile 4th Ur Personality 3rd Ur Nature 2nd Ur SMS and 1st \Ur Lovely Friendship\"... good morning dear"
Ok.

- Here we can see that the messages which has greater length are ham messages thus we can add length as on of the feature
- As the spam messages differ on many basis such as their context and pattern are different we can use it as an asset

Ham messages length vs Spam messages length




0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
200
400
600
800
length

SVM Model

- The following code trains and tests a SVM model using sklearn, The gamma value was achieved by playing around with the figure
- We have selected “sigmoid” as the the kernel or model fitting equation
- The Sorted and filtered dataset is splited with using python library of train_test_split.
- I have selected gamma =1 for this model training.
- After fitting the trained model we get an accuracy of 0.9874
- Thus the model has been fitted well with the training dataset.
- Although inorder to assure the data is not over fitted we have calculated accuracy score using test features which gives accuracy of 0.9784

Multinomial Naive Bayes Model

- The following code trains and tests a Naïve Bayes model using sklearn, The alpha value was achieved by playing around with the figure
- The Sorted and filtered dataset is splited with using python library of train_test_split.
- I have selected gamma =0.2 for this model training after a lot of playing it can be done by for loop too.
- After fitting the trained model we get an accuracy of 0.9854
- Thus the model has been fitted well with the training dataset.
- Although inorder to assure the data is not over fitted we have calculated accuracy score using test features which gives accuracy of 0.9961

Conclusion and Future scope:

- Thus we can se from the predictions that Multinomial Naive Bayes Model worked more efficiently than the SVM model.
- Thus such type of filters can be used in email spam and classification
- Further using CNN and RNN we can form model for problems of Computer vision
-

END