SPAM

Natural Language Processing (NLP)

4/10/24

Contents / Agenda

- Data Dictionary
- Business Problem Overview and Solution Approach
- Exploratory Data Analysis
- Model Performance Summary
- Insights & Recommendations
- Appendix

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Dictionary

- **Category**: Contains the labels 'spam' or 'ham' for the corresponding text data
- **Message**: Contains the SMS text data

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

How to use this deck?

- This slide deck serves as a comprehensive template for your project submission
- Within this deck, you will come across various questions that are intended to test your ability to understand data visualizations, discover patterns / insights and postulate hypothesis. Think thoroughly and provide answers to these questions
- You are encouraged to modify this deck as required, by replacing the questions with suitable answers
- Please feel free to incorporate additional points if you deem necessary

Note: The data visualizations you see in this deck are obtained from RapidMiner

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Business Problem Overview and Solution Approach

- Please define the problem
  - Spam is a major issue in Personal Privacy and Security with respect to Technological innovations such as Short Messaging Services (SMS). SMS spam specifically, is the act of sending unsolicited messages containing advertising (not harmful) as well as malicious (harmful) links. As a Data Science Manager at Cyber Solutions, I have been tasked with utilizing Natural Language Processing (NLP) to identify and classify those malicious messages prior to our employees receiving them on their phones.
- Please mention the solution approach / methodology
  - We are working with only 2 columns in our data set, Category and Message. Our objective will be to create a decision system that can correctly filter the SMS messages as safe or unsafe. If the messages are safe, then we will let them get pushed to employee phones. If they are unsafe, we will restrict them at the 'point of entry.' To create this process, we will utilize techniques such as NLP – diving into the morphology of the message context and continue to refine our model based on the training and testing data results.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA

- Provide a brief overview of the raw dataset

Open in Turbo Prep Auto Model Filter (5,572 / 5,572 examples): no_missing_attrib...

Row No.	Category	Message
1	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	ham	Ok lar... Joking wif u oni...
3	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry q...
4	ham	U dun say so early hor... U c already then say...
5	ham	Nah I don't think he goes to usf, he lives around here though
6	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? ...
7	ham	Even my brother is not like to speak with me. They treat me like aids patent.
8	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune ...
9	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To clai...
10	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for ...
11	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough to...
12	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, ...
13	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM...

ExampleSet (5,572 examples, 0 special attributes, 2 regular attributes)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

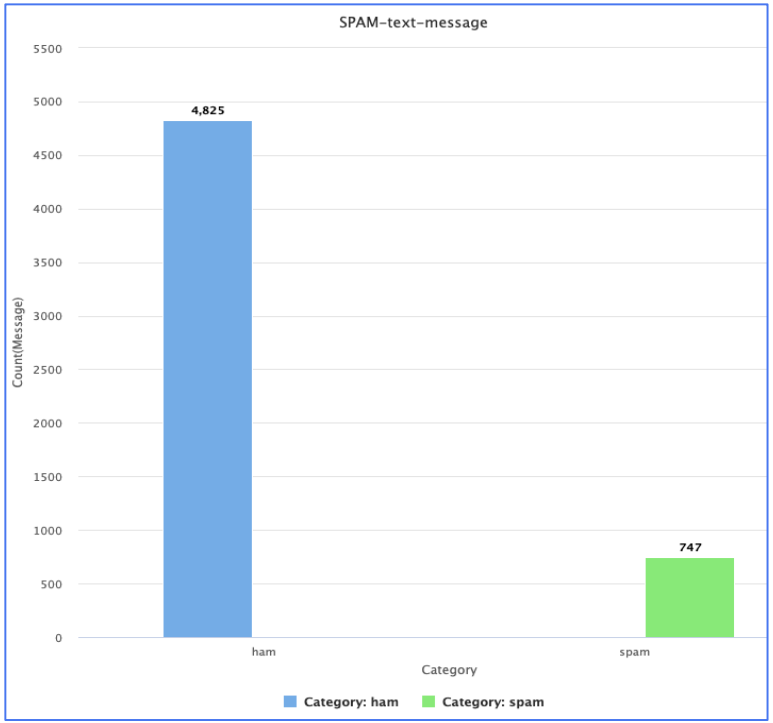
EDA

- Provide a brief overview of the raw dataset
  - The dataset displays 3 columns (Row No., Category, and Message) with 5,572 rows of data
    - Row. No is an automatically assigned digit based on the row
    - Category has two values: Ham for acceptable messages, and Spam for malicious messages
    - Message contains free-text which is where we will look to determine if something is Ham or Spam
  - Of the 5,572 rows of data:
    - Ham consists of 747 rows, or 13.4% of the total dataset
    - Spam consist of 4825 rows, or 86.6% of the total dataset
    - There are no blank values, indicating a complete dataset

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA

- What is the percentage of ham and spam messages in the dataset? How can we handle the imbalance in the dataset?
  - Ham consists of 747 rows, or 13.4% of the total dataset
  - Spam consist of 4825 rows, or 86.6% of the total dataset
  - There are no blank values, indicating a complete dataset
  - Given the large difference in Ham and Spam, we should first look to the Spam dataset to make sure that it is correctly filtering and withholding malicious emails, while sending approved messages



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA: Word Cloud for Spam messages

The below illustration is commonly used when working with textual data. What is this illustration called? How does it help solve the problem statement? When is it used?



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA: Word Cloud for Spam messages

The below illustration is commonly used when working with textual data. What is this illustration called? How does it help solve the problem statement? When is it used?

- The illustration is a Word Cloud which is normally used to depict the words in a particular set of text or dataset. In this specific example, the words are in an unstructured format and appear to all be the same size, indicating they all have equal weight or are unique.

- This image can help us because it is providing a consolidated view of all the words available in one central place. As a result, we can now start to pick off words one by one and put them in either the Spam or Ham bucket, respectively.

- Word Clouds are helpful because they display all of the text to us, and by using lemmatization, can normalize all words to their root form which helps us remove Stop Words as well. We can also do a count on the frequency of certain words to determine their weight in the total population.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Text Analysis

Which text analysis technique is used to find the frequencies of words occurring in documents? How does it help solve a problem statement?

Row No.	word	in documents	total ↓
1	call	480	542
2	get	399	429
3	come	264	283
4	dont	244	270
5	know	238	253
6	free	197	241
7	day	222	238
8	love	199	237
9	time	222	235
10	want	224	235



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Text Analysis

Which text analysis technique is used to find the frequencies of words occurring in documents? How does it help solve a problem statement?

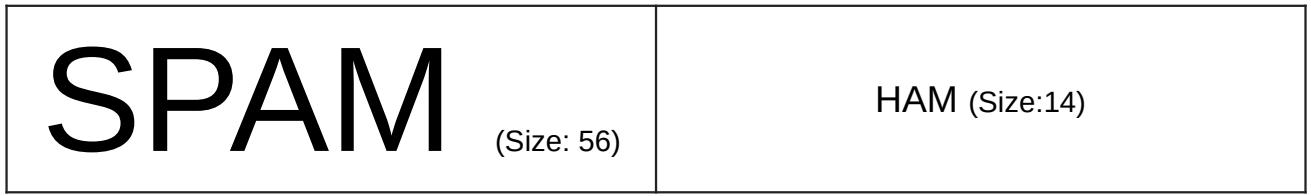
- To better understand the data in this image, we can utilize the Inverse Document Frequency to take our complete, unique set of words and run them against the number of documents (SMS messages) that they appear in.
- Using this information, we can then take the log of these fractions (i.e. the total number of documents/the number of documents containing a word).
- Once we have the Term Frequency and the Inverse Document Frequency for each unique word, we can calculate the weight by multiplying them together. This product will indicate to us, which words to pay more attention to first.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Text Analysis

Which text analysis technique is used to find the frequencies of words occurring in documents? How does it help solve a problem statement?

- To better understand the data in this image, we can utilize Clustering to group all the words together and visually depict which words appear the most.
- This is feasible because we can set the weight of a certain word to the number of times it appears in a dataset (e.g. if a dataset contained 100 words – either Ham: 20 or Spam: 80, Spam appears 4x more than Ham, so it would appear to be 4x bigger – example below.
- Understanding this data is helpful to our analysis because we have proof that certain words appear at a higher frequency in malicious emails as opposed to approved emails, which will help us train our model to flag those words (in this specific case: Call and Get).



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Text Analysis

What is the technique used to determine the sentiment of a piece of textual data? How can positive and negative scores be used to determine the sentiment? How is it helpful in solving business problems?

Row No.	Score	Message	Category
1	0.173	Go until jurong point crazy Available only in bugis n great world la e bu...	ham
2	0.121	Ok lar Joking wif u oni	ham
3	0.243	Free entry in a wkly comp to win FA Cup final tkts st May Text FA to t...	spam
4	0.173	U dun say so early hor U c already then say	ham
5	0.295	Nah I don t think he goes to usf he lives around here though	ham
6	0.156	FreeMsg Hey there darling it s been week s now and no word back I ...	spam
7	0.763	Even my brother is not like to speak with me They treat me like aids pa...	ham
8	0.156	As per your request Melle Melle Oru Minnaminunginte Nurungu Vetta...	ham
9	0.173	WINNER As a valued network customer you have been selected to rec...	spam
10	-0.139	Had your mobile months or more U R entitled to Update to the latest ...	spam

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Text Analysis

What is the technique used to determine the sentiment of a piece of textual data? How can positive and negative scores be used to determine the sentiment? How is it helpful in solving business problems?

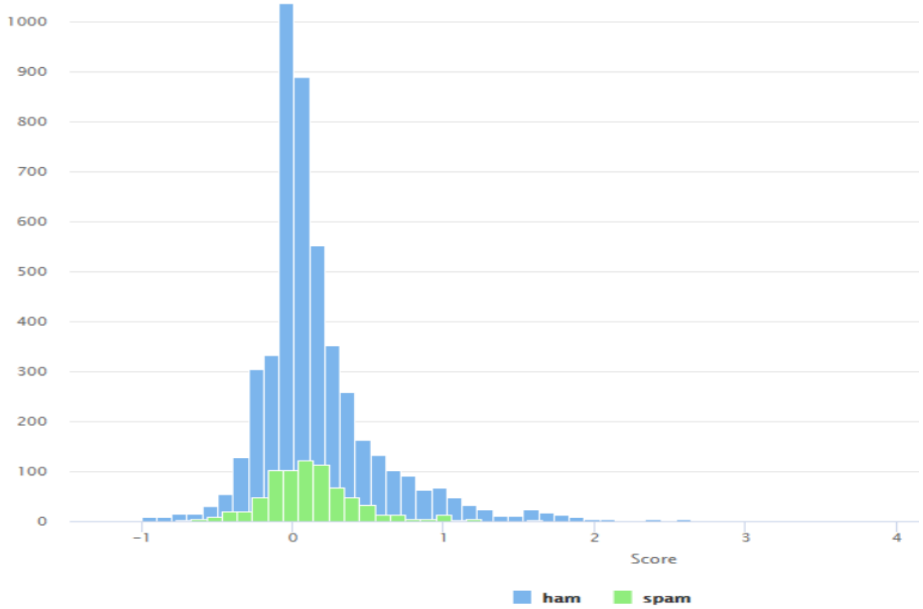
- The better understand the sentiment of textual data, we can do a sentiment analysis to understand whether certain words contain more malice than others.
- In this example, we can see that the Spam messages all contained words such as: Free, (Win)ner, and entitled, whereas Ham messages did not contain those specific words.
- However, we can see that our model has given a high Sentiment score to rows 3 and 5, indicating that we need to refine our model a bit further. Analyzing the data further, it can be inferred that a significant amount of Ham messages have relatively equal Sentiment scores to Spam messages.
- To solve our problem better, we now know that we will need to refine our model further to better understand and classify what is Spam vs. Ham.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Sentiment Scores for Spam and Ham texts

- From a business perspective, why is it important to analyze the sentiment scores of both spam and ham messages in SMS communication? How does the difference in sentiment scores between spam and ham messages provide valuable insights for preventing cyber attacks through SMS?

- It is important to understand the Sentiment scores for both Spam and Ham messages because it will help us refine our model with respect to how we are filtering certain words.
- In the chart to the right, we can see a normal distribution of our entire dataset.
- Ham has a much sharper slope towards zero, from both sides. Whereas Spam does not have as big of a distribution and with a tamer slope.
- This tells us that Ham has a greater concentration of words around the mean compared Spam.
- This tells us that our model has too many words that it is accepting and needs refinement.



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Evaluation Decision Tree

- Please comment on the model performance of the Decision Tree Model
  - Overall, the model seems to be performing well
  - This is because the Training Accuracy, Recall, and Precision are 90%+, and the Testing Accuracy, Recall, and Precision are very close to the training percentages (all 90%+)
  - We have an extremely high accuracy (95%+) for both our Training and Testing data, so we can say that our model is doing a great job at correctly predicting Spam and Ham outcomes
  - This is bolstered with a high precision – both Training and Testing are around 95% indicating that our model is often correct with its positive predictions – correct with Spam and Ham
  - While our Recall is high (90%+), they are lower than the Training and Accuracy, pointing to an area that can be improved. But our model is doing a good job at finding the instances of Spam and Ham

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	97.22	96.59	92.03	91.50	95.71	93.50

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Evaluation Pruned Decision Tree

From the company's perspective, how do the performance metrics of the decision tree model, both before and after pruning, impact the effectiveness of the SMS spam detection system?

- After Pruning our model, we can see that its performance has decreased overall
- Our Training and Testing Accuracy have both fallen approximately 1.5-2%
- Our Training and Testing Recall have both fallen 0.25-.5%
- Our Training and Testing Precision have both fallen, Training – 1.75%; Testing – 5%+

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree Before Pruning	97.22	96.59	92.03	91.50	95.71	93.50
Decision Tree - Pruned	96.75	94.79	91.76	89.90	93.96	88.18

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Evaluation Random Forest

- Please comment on the model performance of the Random Forest Model
  - The Random Forest is working very well with both the Training and Testing Accuracy and Precision, with both 95%+
    - However, our Recall is much lower in the mid 80%
  - This means that our Random Forest model is doing a great job at correctly predicting Spam and Ham outcomes, and that it is correct with its positive predictions
    - However, we need to improve the Recall for both our Training and Testing data, as our model is not doing a good job at finding those positive instances

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest	95.83	95.42	84.45	82.89	97.70	97.49

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Evaluation Pruned Random Forest

Based on the provided evaluation metrics, how does the performance of the random forest model compare to the pruned random forest model in terms of accuracy, recall, and precision for SMS spam detection?

- In the Pruned Random Forest, we see relatively the same performance in the Training Accuracy and Precision, but an improvement in the Training Recall
- However, although we have improved our Testing Recall, our Testing Accuracy has slightly fallen (0.75%) and our Testing Precision has fallen nearly 7%!
- As a result, this indicates that the Pruned Random Forest is overall not a better model than the Random Forest

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest before pruning	95.83	95.42	84.45	82.89	97.70	97.49
Random Forest - Pruned	97.49	94.70	91.70	85.31	97.32	90.78

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- Description of the ML model that best fits the company's objective. Also state which evaluation metric (such as accuracy, recall or precision) is important to achieve the business objective and why?
- Offer suggestions and advice for the company based on the gathered insights.

Model	**Train Accuracy**	**Test Accuracy**	**Train Recall**	**Test Recall**	**Train Precision**	Test Precision
Decision Tree before pruning	97.22	96.59	92.03	91.50	95.71	93.50
Decision Tree - Pruned	96.75	94.79	91.76	89.90	93.96	88.18
Random Forest before pruning	95.83	95.42	84.45	82.89	97.70	97.49
Random Forest - Pruned	97.49	94.70	91.70	85.31	97.32	90.78

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- Description of the ML model that best fits the company's objective. Also state which evaluation metric (such as accuracy, recall or precision) is important to achieve the business objective and why?
- Offer suggestions and advice for the company based on the gathered insights

  - My recommendation to the company is to use the unpruned Decision Tree and work on refining it further to improve the Recall
  - It is our best model because it has higher percentages across the board for Accuracy and Recall
  - While the Precision is not as high as the Random Forest, it is still high (94-95%)
  - Evaluating all these variables holistically, the drop-off with the Random Forest is acceptable because it is much smaller than the drop-offs noticed in the Testing Recall data

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Insights and Recommendations

- Please mention actionable insights & recommendations
  - The unpruned Decision Tree is our best model to use and refine
  - To improve it, we can focus on improving the overall Recall, especially in the Testing dataset
  - We can do this by tracking our model's performance overtime, and by taking away the White Noise, we can understand our data objectively without the trends
  - We can also use more refined clustering techniques to categorize our data more effectively

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Insights and Recommendations

- How can sentiment analysis be utilized to address diverse problem statements? In what ways can it be applied across various industries?
  - Sentiment analysis is extremely helpful in understanding text data because we can better understand if a message consists of malice or is a positive and more acceptable message
  - This is especially critical in the Consumer industry because consumers should have a positive reaction to what is being advertised to them – this is what ultimately drives sales
  - In Healthcare, Sentiment analysis is essential because an illness or disease needs to be carefully communicated to a patient so that they do not lose their mind – this is the ethical part of AI/ML

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Insights and Recommendations

- Select an industry of your choice and pinpoint a prospective issue that could be resolved through text processing. Explain the approach to solving this problem statement.
  - A major issue that can be resolved via text processing is in the Advertisement space. Amazon is already aggregating all customer reviews into one consolidated review, which provides a quick snapshot to the buyer
  - Likewise, at my current company, Southern Glazers Wine & Spirits, we are the largest liquor and wine distributor in the world. We can utilize text processing to understand how customers felt about certain wines, and make better recommendations of wine to the customers so they can improve sales
  - As an example:
    - Store A buys 10 cases of Pinot Grigio (white) and 10 cases of Merlot (Red)
    - From the tastings they offer in their store, they provide feedback that: *the Merlot has the right level of dryness, but it too fruity. Whereas the Pinot Grigio is too sweet.*
    - *From analyzing these types of comments, we can build profiles for the Pinot Grigio and the Merlot so that we communicate to future clients that customers felt the Merlot was fruity, and the Pinot was sweet.*

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

APPENDIX

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

greatlearning
Power Ahead

Happy Learning !

