MSBD5018 Natural Language Processing

Automated Contextual Spam Categorization

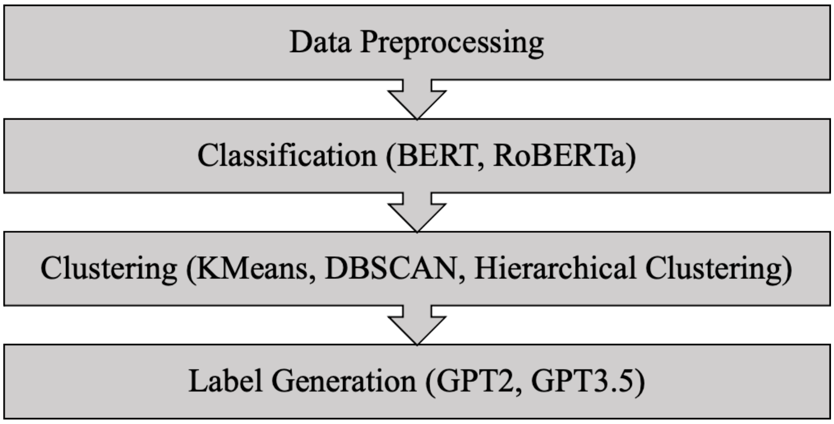
Group 20

LAKHANI Sunil Harsh
20910249

LI Ka Ho
20922151

TSANG Kai Ho
20905476

Project Flow


Data Preprocessing
Classification (BERT, RoBERTa)
Clustering (KMeans, DBSCAN, Hierarchical Clustering)
Label Generation (GPT2, GPT3.5)

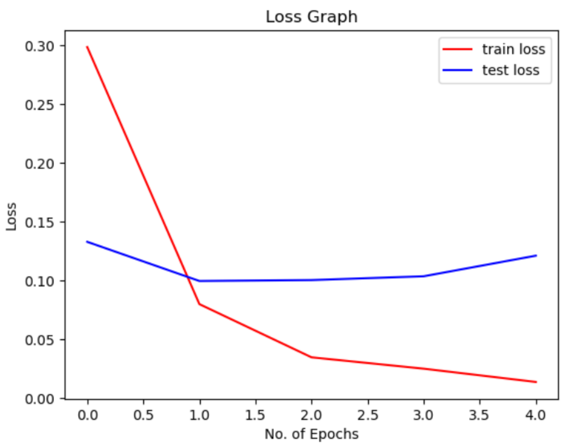
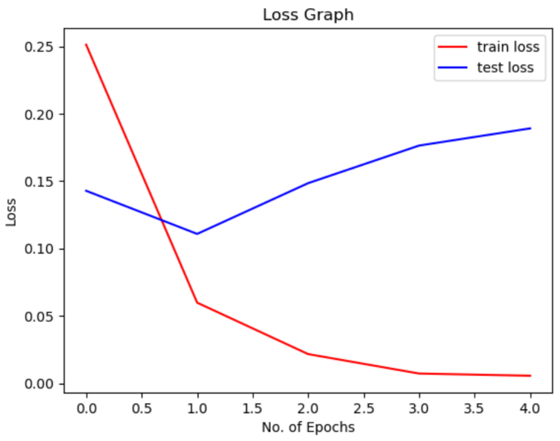
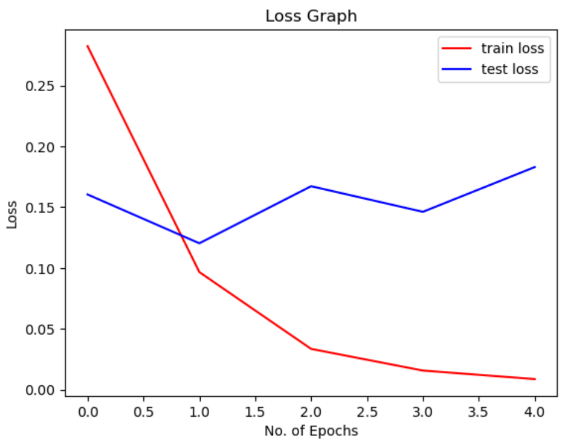
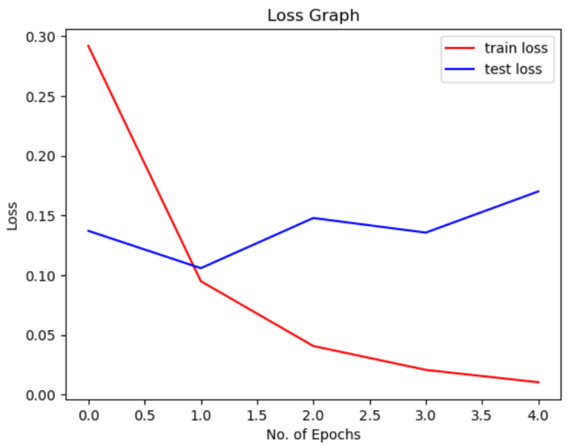
Classification

To detect whether an input is spam or not, we utilized two language models, namely BERT and RoBERTa, to train a classifier. We fine-tune the models with extra layers (named fine-tuned) and without extra layers (named base) with 5 epochs.

	BERT (base)	BERT (fine-tuned)	RoBERTa (base)	RoBERTa (fine-tuned)
Train Loss	0.0802	0.0598	0.0158	0.0208
Validation Loss	0.0999	0.1108	0.1462	0.1357
Validation Accuracy	0.9703	0.9661	0.9703	0.9686
F-1 Score	0.9703	0.9661	0.9700	0.9661
Recall	0.9703	0.9661	0.9700	0.9661
Precision	0.9704	0.9663	0.9703	0.9663

** Values are rounded to 4 decimal places*

Loss Curves:

BERT (base)	BERT (fine-tuned)
	
RoBERTa (base)	RoBERTa (fine-tuned)
	

The results from our experiments on training spam classifiers showed that each model we trained experienced a drastic drop in training loss during the first epoch, followed by a steady decrease in loss until it reached a plateau. We trained each model for 5 epochs to evaluate their performance and compare them against each other.

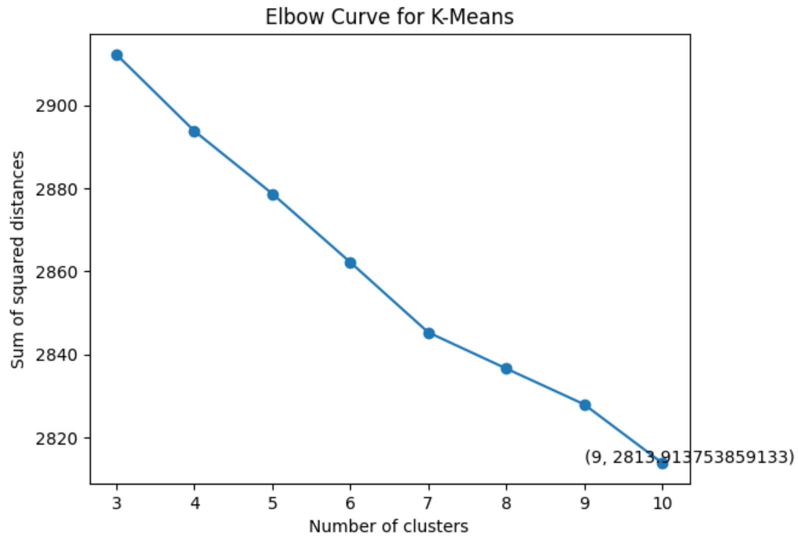
From the results table and loss curves, it is clear that the best performing model is the BERT (base) model, with an impressive accuracy of almost 97%. Although the other models we trained were not far behind, we ultimately decided to use BERT (base) as our spam classifier in the final pipeline.

The BERT (base) model outperformed other models and fine-tuning experiments we conducted, suggesting that it is highly effective at detecting spam content. Its success can be attributed to its advanced natural language processing capabilities, which allows it to understand the meaning and context of the content accurately. Overall, we are confident that the BERT (base) model will be an effective spam classifier for the final pipeline.

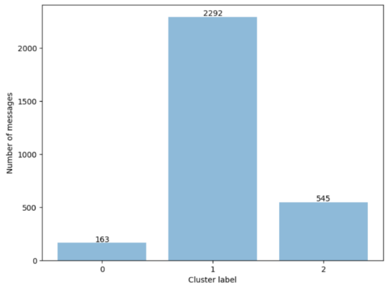
Clustering

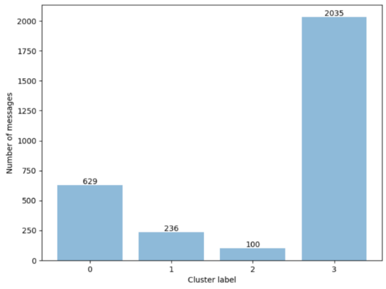
To investigate the categories of spam content, we first utilized three clustering algorithms to identify clusters of common types of spam content

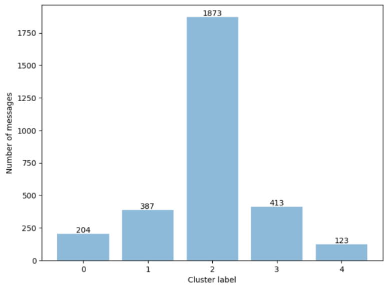
K-means:



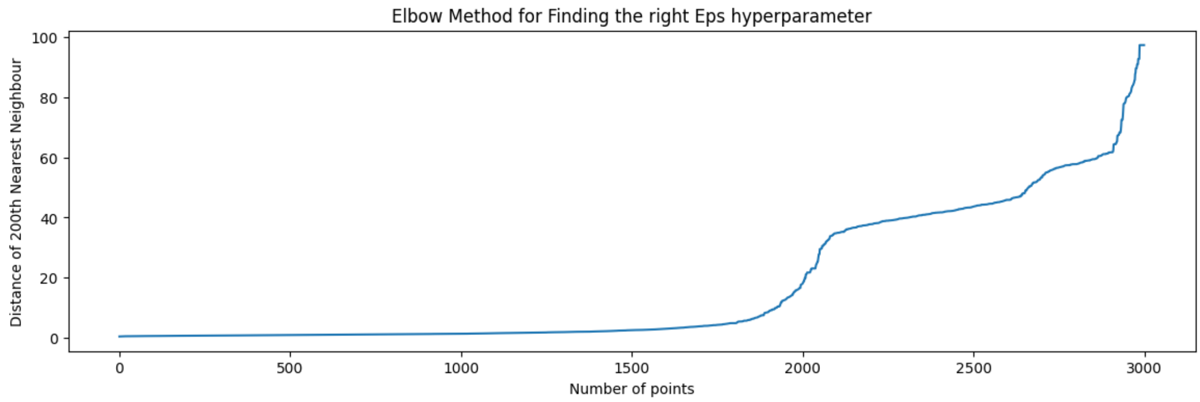
The figure above shows that the additional cost is no longer justified by the diminishing returns after seven clusters. This means the performance of our model will not benefit much from incorporating another cluster after seven clusters. However, we believed that categorizing spam content into seven clusters is slightly excessive, which led us to experiment with three, four, and five clusters with generative models. The optimal number of clusters will be determined by our personal judgement on the generated labels.



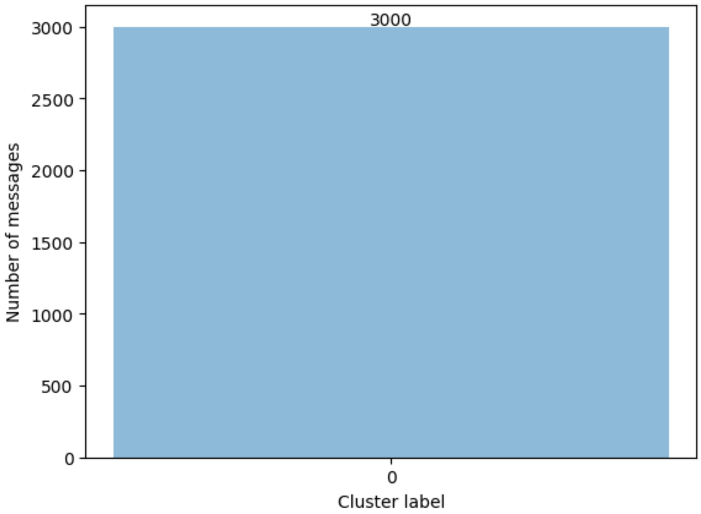




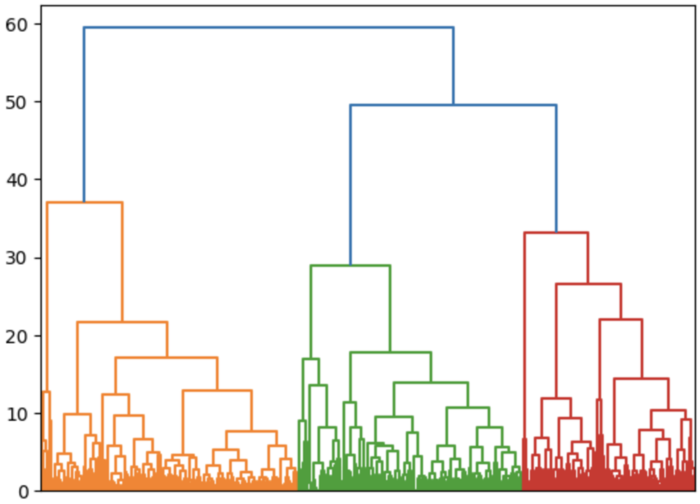
DBSCAN:



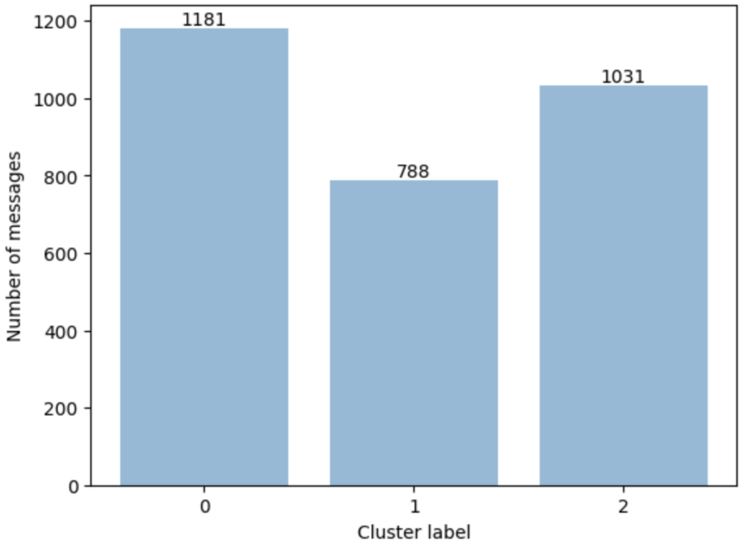
Unlike K-means, DBSCAN does not require the number of clusters to be pre-defined. Instead, it identifies clusters based on the density of data points and can also identify noise or outliers. From the above figure, we selected an epsilon value of 40 based on the knee point and trained the model accordingly. The DBSCAN algorithm suggested the presence of a single cluster as shown below.



Hierarchical:



By running the Hierarchical clustering algorithm, we generated a dendrogram that suggests splitting the spam dataset into 3 clusters, determined by the distance among the clusters. The distribution of number of messages within each cluster can be seen down below.



We will experiment and compare each clustering algorithm with the generative algorithm in the label generation section down below.

Label Generation

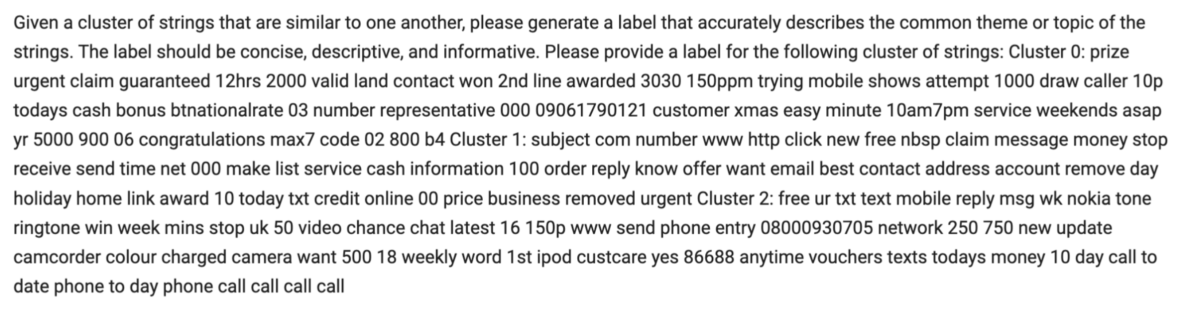
To determine the most suitable clustering model for our spam dataset, we employed two generative models and designed a prompt to generate a label for each cluster. As the k-means algorithm was insufficient in identifying the optimal number of clusters, together with hierarchical clustering, we experimented with four clustering models. While we initially tested GPT-2 on various numbers of k-means clusters, the results were unsatisfactory, prompting us to exclusively use ChatGPT-3.5 for comparison with hierarchical clustering.

Ultimately, we chose the optimal clustering model for our spam dataset through a voting system based on whether the labels accurately represented the clusters. Moreover, after experimenting with all of the clustering algorithms, we came to a conclusion that the k-means clustering results were more balanced which is why we selected it to be a part of the final pipeline.

GPT-2:

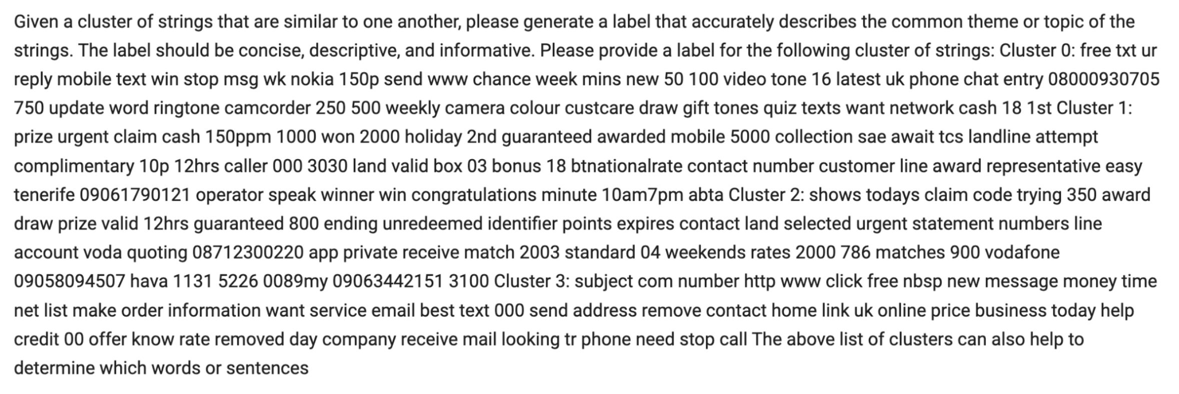
For each number of clusters we found using the k-means algorithm, we made use of GPT-2 to generate labels based on the prompt and response below:

Number of clusters = 3



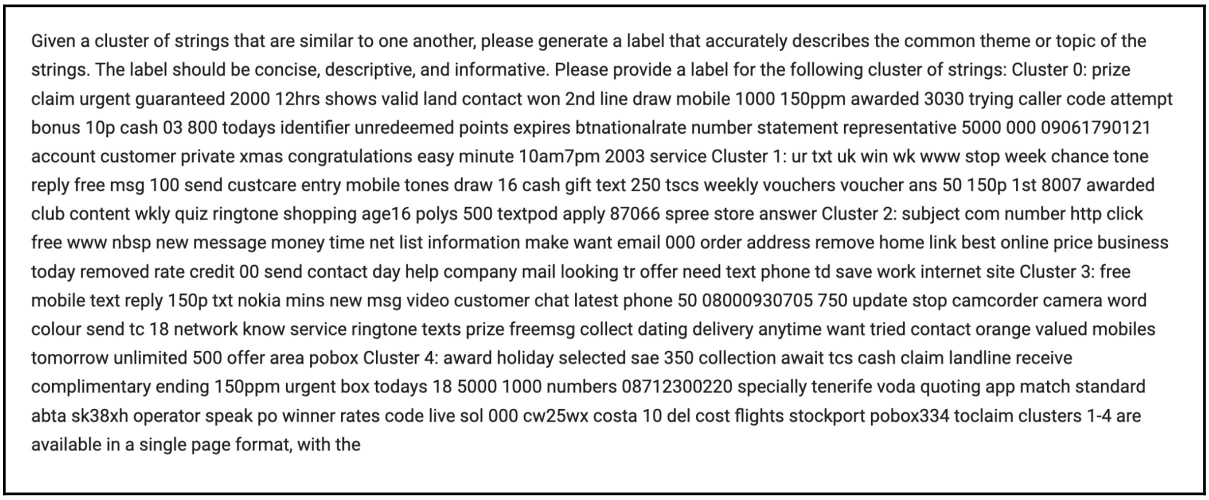
Given a cluster of strings that are similar to one another, please generate a label that accurately describes the common theme or topic of the strings. The label should be concise, descriptive, and informative. Please provide a label for the following cluster of strings: Cluster 0: prize urgent claim guaranteed 12hrs 2000 valid land contact won 2nd line awarded 3030 150ppm trying mobile shows attempt 1000 draw caller 10p todays cash bonus btnationalrate 03 number representative 000 09061790121 customer xmas easy minute 10am7pm service weekends asap yr 5000 900 06 congratulations max7 code 02 800 b4 Cluster 1: subject com number www http click new free nbsp claim message money stop receive send time net 000 make list service cash information 100 order reply know offer want email best contact address account remove day holiday home link award 10 today txt credit online 00 price business removed urgent Cluster 2: free ur txt text mobile reply msg wk nokia tone ringtone win week mins stop uk 50 video chance chat latest 16 150p www send phone entry 08000930705 network 250 750 new update camcorder colour charged camera want 500 18 weekly word 1st ipod custcare yes 86688 anytime vouchers texts todays money 10 day call to date phone to day phone call call call call

Number of clusters = 4



Given a cluster of strings that are similar to one another, please generate a label that accurately describes the common theme or topic of the strings. The label should be concise, descriptive, and informative. Please provide a label for the following cluster of strings: Cluster 0: free txt ur reply mobile text win stop msg wk nokia 150p send www chance week mins new 50 100 video tone 16 latest uk phone chat entry 08000930705 750 update word ringtone camcorder 250 500 weekly camera colour custcare draw gift tones quiz texts want network cash 18 1st Cluster 1: prize urgent claim cash 150ppm 1000 won 2000 holiday 2nd guaranteed awarded mobile 5000 collection sae await tcs landline attempt complimentary 10p 12hrs caller 000 3030 land valid box 03 bonus 18 btnationalrate contact number customer line award representative easy tenerife 09061790121 operator speak winner win congratulations minute 10am7pm abta Cluster 2: shows todays claim code trying 350 award draw prize valid 12hrs guaranteed 800 ending unredeemed identifier points expires contact land selected urgent statement numbers line account voda quoting 08712300220 app private receive match 2003 standard 04 weekends rates 2000 786 matches 900 vodafone 09058094507 hava 1131 5226 0089my 09063442151 3100 Cluster 3: subject com number http www click free nbsp new message money time net list make order information want service email best text 000 send address remove contact home link uk online price business today help credit 00 offer know rate removed day company receive mail looking tr phone need stop call The above list of clusters can also help to determine which words or sentences

Number of clusters = 5



Given a cluster of strings that are similar to one another, please generate a label that accurately describes the common theme or topic of the strings. The label should be concise, descriptive, and informative. Please provide a label for the following cluster of strings: Cluster 0: prize claim urgent guaranteed 2000 12hrs shows valid land contact won 2nd line draw mobile 1000 150ppm awarded 3030 trying caller code attempt bonus 10p cash 03 800 todays identifier unredeemed points expires btnationalrate number statement representative 5000 000 09061790121 account customer private xmas congratulations easy minute 10am7pm 2003 service Cluster 1: ur txt uk win wk www stop week chance tone reply free msg 100 send custcare entry mobile tones draw 16 cash gift text 250 tscs weekly vouchers voucher ans 50 150p 1st 8007 awarded club content wkly quiz ringtone shopping age16 polys 500 textpod apply 87066 spree store answer Cluster 2: subject com number http click free www nbsp new message money time net list information make want email 000 order address remove home link best online price business today removed rate credit 00 send contact day help company mail looking tr offer need text phone td save work internet site Cluster 3: free mobile text reply 150p txt nokia mins new msg video customer chat latest phone 50 08000930705 750 update stop camcorder camera word colour send tc 18 network know service ringtone texts prize freemsg collect dating delivery anytime want tried contact orange valued mobiles tomorrow unlimited 500 offer area pobox Cluster 4: award holiday selected sae 350 collection await tcs cash claim landline receive complimentary ending 150ppm urgent box todays 18 5000 1000 numbers 08712300220 specially tenerife voda quoting app match standard abta sk38xh operator speak po winner rates code live sol 000 cw25wx costa 10 del cost flights stockport pobox334 toclaim clusters 1-4 are available in a single page format, with the

After conducting the experiments, we discovered that GPT-2 has a limited maximum text length, allowing only up to 50 words per cluster to be included in the prompt. This constraint may hinder the generalizability of the clusters. In addition, we encountered issues with GPT-2's interpretation of the prompt. Rather than generating labels for the clusters, it either continues writing the prompt or provides a summary of the entire prompt.

ChatGPT-3.5:

For each number of clusters we found using the k-means algorithm, GPT-3.5 generated responses as follows:

Number of clusters = 3
Cluster 0: Prize/Winning Claims and Offers Cluster 1: Marketing Lists and Communication Cluster 2: Mobile Phone Promotions and Services
Number of clusters = 4
Cluster 0: Mobile phone promotions and services Cluster 1: Winning prizes and offers Cluster 2: Information, numbers, and codes Cluster 3: Email and online marketing and services
Number of clusters = 5
Cluster 0: Prize Claim and Winning Notification Cluster 1: Mobile Text and Tone Services Cluster 2: Email Marketing and Information Lists Cluster 3: Holiday and Travel Offers Cluster 4: Award and Complimentary Collection

For clusters we found using the hierarchical clustering algorithm, ChatGPT-3.5 generated responses as follows:

Number of clusters = 3

```
Cluster 0: Business and Financial Information
Cluster 1: Email Marketing and Communication
Cluster 2: Sales and Promotions
```

Conclusion:

We are pleased to announce that we have successfully integrated our trained spam classifier, clustering algorithm, and generative model to create an automated contextual spam categorization tool. This tool takes an input, classifies it as either spam or legitimate, and if it is spam, categorizes it into the appropriate category.

To test the efficacy of our tool, we created customized spam inputs and manually tested the model. We found that the model performed exceptionally well, correctly categorizing spam content and clustering them into the appropriate category with high accuracy.

To further validate our model, we tested it on the YouTube comments dataset. The dataset is challenging to classify since it contains a wide range of content, some of which are not strictly spam but could be considered as such. Nevertheless, our tool achieved an accuracy of approximately 77% of correctly categorizing the content as spam. We consider this a significant achievement, given the complexity of the dataset and the challenges it poses.

Our spam categorization tool is a significant step forward in automated spam detection and filtering. It has enormous potential in reducing the number of unwanted spam content that users come across daily, particularly in environments where spam content can be particularly prevalent. Additionally, our tool can be easily customized to suit specific contexts and spam content types, making it a versatile solution that can be applied to a wide range of use cases.

In conclusion, we are confident that our automated contextual spam categorization tool will have a significant impact on reducing the amount of spam content users come across, improving their experience and reducing the risk of phishing attacks. We look forward to seeing its impact in various contexts and industries.