SCSUG 2017

Classifying and Predicting Spam Messages using Text Mining in SAS® Enterprise Miner™

Mounika Kondamudi, Oklahoma State University, Mentored by Balamurugan Mohan, H&R Block

Abstract

In this technologically advanced digital world, identifying a spam message is of extreme importance. Spam messages are generally unsolicited and unwanted messages and when accessed can trap people in scam subscriptions that might infect their devices with malicious software. Sometimes this can be even more annoying to the recipient because, unlike in email, some recipients may be charged a fee for every message received.

The dataset used for this analysis is a collection of 6,927 Spam and Ham (General conversation, anti-spam) messages that include 5,574 (747 spam, 4,827 ham) English messages from UCI Machine Learning Repository and a corpus of 1,353 spam messages from Dublin Institute of Technology (DIT). This paper motivates work on identifying clusters of high frequency spam and ham words. A classification model, which can classify and predict the messages as spam and ham based on the rules built by the text builder node, is discussed. The predictive power of this model is assessed by the misclassification rate in the scored data (5%).

Introduction

Mobile or SMS spam is a growing problem particularly because of the availability of bulk pre-pay SMS packages and also because of the fact that there is a higher response rate as it is a trusted and more personal service. Many android apps are available to block spam texts and mobile carriers, too offer various spam-blocking services. That being said, there are always some spam texts that will get through and spammers will do their best to escape antispam technology. Using text mining we can find the terms that are most commonly used in a spam message. We can analyze each term and also see how strongly it is associated with other text terms. We can also identify the set of rules based on which the messages can be classified as either spam or ham using the content categorization code. These rules help in predicting the category of a message.

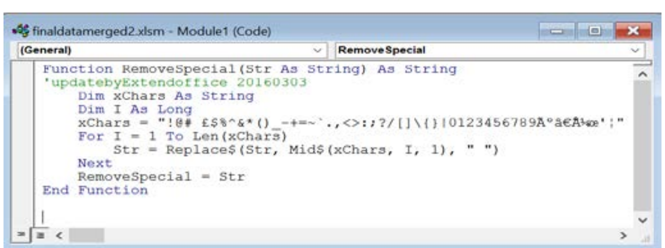
A working model of this, when implemented successfully, can be very helpful to both customers and companies. Carrier companies can protect their customers from spammers and their spam texts. Companies can use the list of high frequency spam words and take necessary precautions to not include these words in their promotional offers.

Data Dictionary

Dataset used for analysis is a collection of 5,574 (747 spam messages and 4,827 ham messages) English messages from UCI Machine Learning Repository and a corpus of 1,353 unique spam messages from Dublin Institute of Technology.

Data Preparation and Cleaning

The messages are initially cleaned for special and unidentifiable characters using the below VBA code. These special and unidentifiable characters are later on added to the stop list in the text parsing node.




```
Function RemoveSpecial(Str As String) As String
'updatebyExtendoffice 20160303
    Dim xChars As String
    Dim I As Long
    xChars = "!@# £$%^&*()_-+=~`.,<>:;?/[]\{}|0123456789Â°â€Â¼œ';"
    For I = 1 To Len(xChars)
        Str = Replace$(Str, Mid$(xChars, I, 1), " ")
    Next
    RemoveSpecial = Str
End Function
```

Fig1: VBA macro snippet

Data Dictionary

Variable	Level	Description
ID	ID	This field represents the unique message number.
Text	Text	This field represents the actual message which is either spam or ham.
Target	Target	This field represents the actual category of the message.

Fig2: Data Dictionary

Methodology

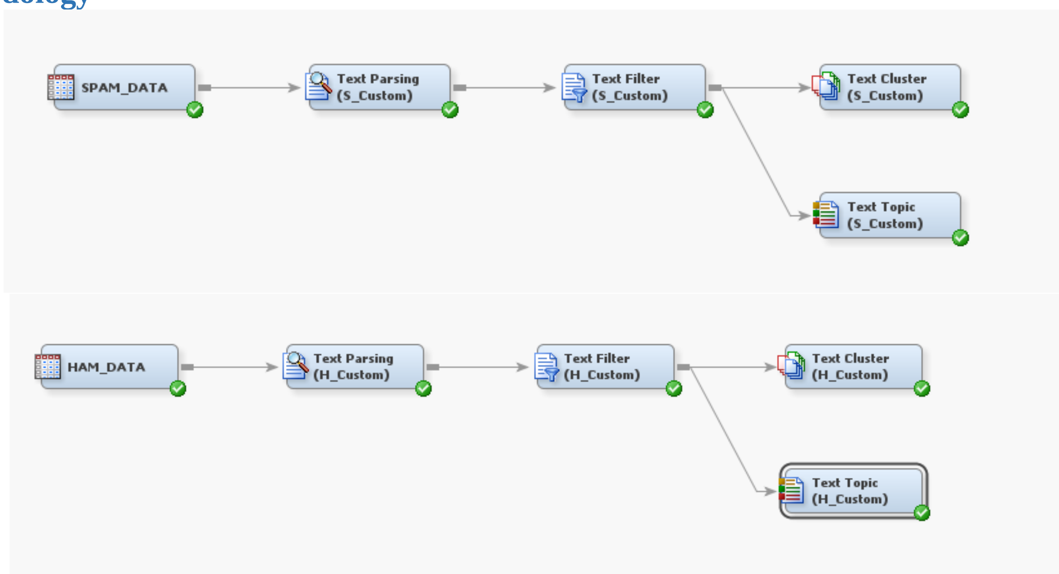


Fig3: To generate and summarize topics from spam and ham messages, as well as classify messages into spam and ham groups.

The data sets used for this analysis are:

- spam_data.sas7bdat
- ham_data.sas7bdat
- spam_stopwords_manual.sas7bdat
- engdict.sas7bdat
- spam_syn_manual.sas7bdat
- syn_py1_dropped.sas7bat

Datasets

Since the data is available as a single SAS file, for the purpose of this analysis, the data set has been divided into spam and ham data. These 2 data sets are added as input sources in SAS Enterprise Miner.

Text Parsing

The text parsing node is connected to the data and a few modifications are made to clean the text data. Using the properties panel,

- The ‘detect different parts of speech’ option is set to ‘no’ to be able to treat the same words or terms with different parts of speech as same terms.
- ‘Num’, ‘Prop’, ‘Verbadj’ parts of speech have been ignored apart from the default options.
- A customized stop words list is identified in order to mask special characters, meaningless words along with the default set of stop words provided by SAS.

The text parsing node generated a term by document matrix which can be used to identify the most frequently occurring words and the number of documents each word has occurred in.

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ text		Alpha	627	524	Y	+	261	2
free		Alpha	598	489	Y		32	5
+ claim		Alpha	538	458	Y	+	262	7
+ reply		Alpha	484	449	Y	+	209	8
+ message		Alpha	432	393	Y	+	2086	10
+ win		Alpha	319	299	Y	+	362	12
+ contact		Alpha	223	223	Y	+	809	14
+ pound		Alpha	220	220	Y	+	2820	15
+ service		Alpha	222	215	Y	+	1692	16
+ prize		Alpha	238	210	Y	+	2209	19
+ minute		Alpha	213	199	Y	+	1265	21
stop		Alpha	234	197	Y		124	23
accident		Alpha	194	194	Y		952	24
+ world wide web		Alpha	192	192	Y	+	1462	25
+ text stop	Noun Group	Alpha	187	187	Y	+	1418	27
+ entitle		Alpha	184	184	Y	+	54	28
+ record		Alpha	175	175	Y	+	2388	29
urgent		Alpha	176	174	Y		3177	30
+ number		Alpha	184	173	Y	+	544	33
+ week		Alpha	209	171	Y	+	45	34
+ phone		Alpha	171	159	Y	+	3087	36
+ indicate		Alpha	154	154	Y	+	2331	37
+ UK		Alpha	152	144	Y	+	2928	38
+ po box		Alpha	141	141	Y	+	2596	39
opt		Alpha	141	140	Y		1852	40
+ guarantee		Alpha	138	138	Y	+	108	41
cash		Alpha	139	136	Y		877	42
+ mobile		Alpha	142	128	Y	+	1856	45
+ line		Alpha	127	127	Y	+	1898	46
+ your mobile		Alpha	127	127	Y	+	1780	46
+ terms and conditions		Alpha	126	126	Y	+	1517	48
yes		Alpha	124	124	Y		574	49
find out		Alpha	118	118	Y		1619	52
free reply	Noun Group	Alpha	118	118	Y		2426	52
+ award		Alpha	128	112	Y	+	2711	56
+ tone		Alpha	174	111	Y	+	2261	57
+ land		Alpha	110	110	Y	+	2612	58
+ customer		Alpha	109	109	Y	+	2117	59
nokia		Alpha	139	109	Y		338	59
+ date		Alpha	116	108	Y	+	1570	61
+ show		Alpha	108	108	Y	+	2863	61
+ receive		Alpha	110	103	Y	+	1538	65
+ draw		Alpha	105	102	Y	+	1670	66
freemsg		Alpha	102	102	Y		1971	66
+ know		Alpha	128	100	Y	+	3023	68
+ want		Alpha	106	98	Y	+	3036	69
compensation		Alpha	97	97	Y		1018	70
+ landline		Alpha	84	84	Y	+	686	75
+ date service	Noun Group	Alpha	80	80	Y	+	2059	76
+ offer		Alpha	85	80	Y	+	324	76
+ sms		Alpha	82	79	Y	+	2786	79
+ day		Alpha	88	76	Y	+	56	81
+ voucher		Alpha	79	76	Y	+	1785	81

Fig4: Text Parsing results for spam data

Some of the most frequent words are text, free, claim, reply, message etc. which makes sense because these are the words commonly used by spammers in their messages. Misspelt words if any are later on removed by the text filter node.

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ good		Alpha	381	342	Y	+	1896	10
+ know		Alpha	295	278	Y	+	7301	15
+ want		Alpha	237	230	Y	+	7340	24
+ love		Alpha	268	221	Y	+	3324	26
+ day		Alpha	241	220	Y	+	134	27
+ don't		Mixed	220	201	Y	+	4632	29
+ late		Alpha	195	190	Y	+	104	30
+ time		Alpha	166	158	Y	+	698	38
+ darling		Alpha	171	155	Y	+	1006	39
home		Alpha	156	154	Y		2792	40
+ night		Alpha	154	144	Y	+	4221	46
today		Alpha	139	138	Y		703	48
+ message		Alpha	140	136	Y	+	4989	49
+ tomorrow		Alpha	140	135	Y	+	4990	50
i am		Alpha	138	134	Y		4049	51
+ back		Alpha	136	131	Y	+	4130	54
dont		Alpha	137	127	Y		5073	55
+ meet		Alpha	131	123	Y	+	866	56
+ work		Alpha	128	122	Y	+	3147	57
+ hope		Alpha	122	116	Y	+	3586	59
+ leave		Alpha	116	113	Y	+	158	63
well		Alpha	106	105	Y		1921	68
+ miss		Alpha	124	103	Y	+	7527	70
+ right		Alpha	106	102	Y	+	6420	71
+ thing		Alpha	111	102	Y	+	4996	71
+ friend		Alpha	108	99	Y	+	1755	73
+ text		Alpha	101	97	Y	+	655	74
+ wait		Alpha	100	97	Y	+	6006	74
+ feel		Alpha	104	95	Y	+	5936	77
+ great		Alpha	101	95	Y	+	4864	77
+ dear		Alpha	104	93	Y	+	1946	81
+ phone		Alpha	93	89	Y	+	7474	82
+ happy		Alpha	103	86	Y	+	1998	84
+ win		Alpha	91	86	Y	+	875	84
yeah		Alpha	86	85	Y		2027	87
+ sleep		Alpha	85	82	Y	+	3160	90
yes		Alpha	82	82	Y		1377	90
+ did not		Alpha	80	80	Y	+	3529	94
lol		Alpha	74	74	Y		1894	98
+ minute		Alpha	75	74	Y	+	2988	98
+ morning		Alpha	82	74	Y	+	972	98
+ babe		Alpha	74	73	Y	+	3864	102
+ week		Alpha	78	73	Y	+	105	102
+ care		Alpha	72	70	Y	+	1753	107
+ buy		Alpha	75	69	Y	+	3752	109
+ number		Alpha	76	69	Y	+	1289	109
+ keep		Alpha	72	68	Y	+	5635	111
+ life		Alpha	81	67	Y	+	2474	113
+ watch		Alpha	69	66	Y	+	7496	115
+ year		Alpha	69	66	Y	+	2096	115
last		Alpha	65	65	Y		7800	117
+ mean		Alpha	65	65	Y	+	7601	117
+ tonight		Alpha	68	65	Y	+	1234	117

Fig5: Text Parsing results for ham data

Text Filter

The text filter node, which is added after the text parsing node, filters out the terms that occurs the least number of times as specified by the user in the properties panel.

- Minimum number of documents is set to 4.
- Text filter node also performs spell check. By enabling this option in the text filter node, synonyms are created for the misspelt words.
- Customized English dictionary is added in the properties panel.
- Customized synonym list is created using python script for all the words that are kept by the text filter node (process described below). This list is imported into the text filter node using the import synonyms ellipsis button in the properties panel.
- Terms to view is changed to 'selected' in the properties panel in order to get a holistic view of the words that are only kept by the text filter node.
- Concept links are identified for some of the most frequent terms using the filter viewer interactive ellipsis button in the properties panel.

Creating synonym list using python

A python script is used to extract the synonyms for the most frequent spam and ham words that were obtained from the text filter node with default properties. PyDictionary is a Dictionary Module for Python to get meanings, translations, synonyms and Antonyms of words. It uses WordNet for getting meanings, Google for translations, and thesaurus.com for getting synonyms and antonyms. PyDictionary module can extract meanings for 250 words at a time and synonyms for a total of 1,418 parent terms were scraped. All these terms were then placed in a document which was later converted into a SAS dataset compatible to be used in Text Filter node as shown in the Fig7. (Shows a partial list of synonyms obtained using python script).

```
from PyDictionary import PyDictionary
dictionary=PyDictionary()


dictionary = PyDictionary("text", ...... ,"free","message")
print (dictionary.getSynonyms())
```

Fig6: Python code snippet

	term	termrole	parent	parentrole
1	good		able	
2	adept		able	
3	capable		able	
4	apt		able	
5	competent		able	
6	welcome		accept	
7	obtain		accept	
8	take		accept	
9	get		accept	
10	acquire		accept	
11	entry		access	
12	connection		access	
13	approach		access	
14	entrance		access	
15	entrée		access	
16	disaster		accident	
17	mishap		accident	
18	calamity		accident	
19	setback		accident	
20	hazard		accident	
21	unwittingly		accidentally	
22	unintentionally		accidentally	
23	haphazardly		accidentally	
24	by mistake		accidentally	
25	fortuitously		accidentally	
26	story		account	
27	explanation		account	
28	detail		account	
29	tale		account	

Fig7: Customized synonym list

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
⊟ reply	490	453	☑	1.0		Alpha
replied	5	5				Alpha
rrply	1	1				Alpha
rpl	2	2				Alpha
rply	17	17				Alpha
replys	2	2				Alpha
replying	11	11				Alpha
reply	448	419				Alpha
repy	1	1				Alpha
replies	3	3				Alpha
⊟ message	432	393	☑	1.0		Alpha
messages	34	31				Alpha
message	142	135				Alpha
messaging	8	8				Alpha
msgs	19	19				Alpha
msg	229	225				Alpha

Fig8: Synonyms grouping

Fig8 from the interactive filter viewer shows synonyms for the words ‘reply’ and ‘message’. Similar terms and misspelt terms are formed into groups using the synonyms that are imported manually and using the English dictionary.

	Parent # Docs	Term	# Docs	Parent	Role	Parent Role	Min Distance	Dictionary	Key	Parent ID
1	489.0	fre	1.0	free			8.0	N	2423.0	32.0
2	4.0	qoute ref	1.0	quote ref	NOUN_GROUP	NOUN_GROUP	10.0	N	2748.0	153.0
3	419.0	repy	1.0	reply			12.0	N	1413.0	209.0
4	419.0	replys	2.0	reply			6.0	N	632.0	209.0
5	18.0	congrat	1.0	congrats			4.0	N	1084.0	290.0
6	27.0	secreat	1.0	secret			8.0	N	1790.0	317.0
7	109.0	nokias	6.0	nokia			6.0	N	1359.0	338.0
8	5.0	recorvery	1.0	recovery			6.0	N	692.0	363.0
9	35.0	unsubscribed	2.0	unsubscribe			2.0	N	1367.0	447.0
10	57.0	landline claim	3.0	land line claim	NOUN_GROUP	NOUN_GROUP	6.0	N	384.0	493.0
11	7.0	voice mail message	1.0	voicemail message	NOUN_GROUP	NOUN_GROUP	4.0	N	93.0	531.0
12	160.0	numberx	1.0	number			5.0	N	1673.0	544.0
13	7.0	superb	2.0	super			12.0	Y	2626.0	669.0
14	5.0	filth	3.0	filthy			12.0	Y	2859.0	703.0
15	41.0	inof	1.0	info			12.0	N	1722.0	797.0
16	179.0	contack	1.0	contact			14.0	N	779.0	809.0
17	179.0	contac	1.0	contact			5.0	N	2645.0	809.0
18	6.0	complementary tenerife	2.0	complimentary tenerife	NOUN_GROUP	NOUN_GROUP	8.0	N	893.0	820.0
19	16.0	vist	6.0	visit			12.0	N	2203.0	947.0
20	194.0	acident	1.0	accident			3.0	N	1060.0	952.0
21	194.0	acceident	1.0	accident			6.0	N	3196.0	952.0
22	194.0	acckent	1.0	accident			12.0	N	321.0	952.0
23	16.0	goverment	1.0	government			5.0	N	1801.0	962.0
24	16.0	givernment	1.0	government			10.0	N	1585.0	962.0
25	16.0	govenment	1.0	government			5.0	N	958.0	962.0
26	97.0	compenation	1.0	compensation			4.0	N	14.0	1018.0
27	25.0	mobi	1.0	mob			10.0	N	3130.0	1073.0
28	65.0	cha	3.0	chat			10.0	N	437.0	1143.0
29	4.0	fil	2.0	film			10.0	N	1502.0	1234.0
30	53.0	wanting	2.0	waiting			14.0	N	3211.0	1284.0
31	10.0	gotto	2.0	goto			6.0	N	438.0	1350.0
32	58.0	recieve	3.0	receive			7.0	N	2743.0	1538.0
33	58.0	receivea	3.0	receive			4.0	N	2063.0	1538.0
34	5.0	sppok	2.0	spook			15.0	N	1845.0	1546.0

Fig9: Text filter spellcheck

Fig9 shows the list of misspelt words in the ‘Term’ column and their corrections in ‘Parent’ column. Text Filter node makes use of the customized English Dictionary that is added in the properties panel.

Concept Links

Concept links can be viewed in the interactive filter viewer from the properties panel of text filter node. It is a type of association analysis between the terms used. They can be created for all the terms that are present in the documents, however it is meaningful to create only for a few important terms. It shows the term to be analyzed in the center and the terms that it is mostly used with as links.

The width of the link depicts the strength of association. The wider the link the stronger is the association and the more important it is. Concept links also show how many times the two terms co-exist together in a sentence.

Concept Links for Spam Data

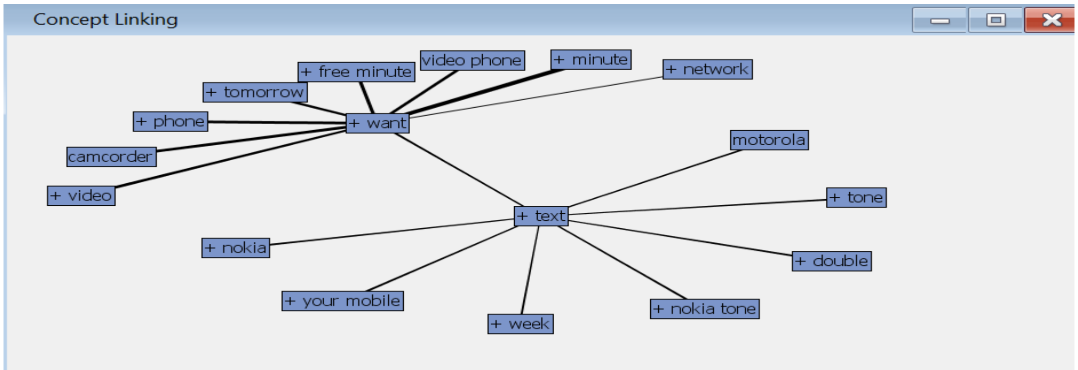


Fig10: Concept link for 'text'

From Fig10, 'Text' is strongly associated with the word 'want'. This means spammers are asking the customers to text back if they want either a free minute or a camcorder or a video phone. The term 'want' is strongly associated with 'minute' and 'camcorder' which means customers are offered free minutes and camcorders if they reply.

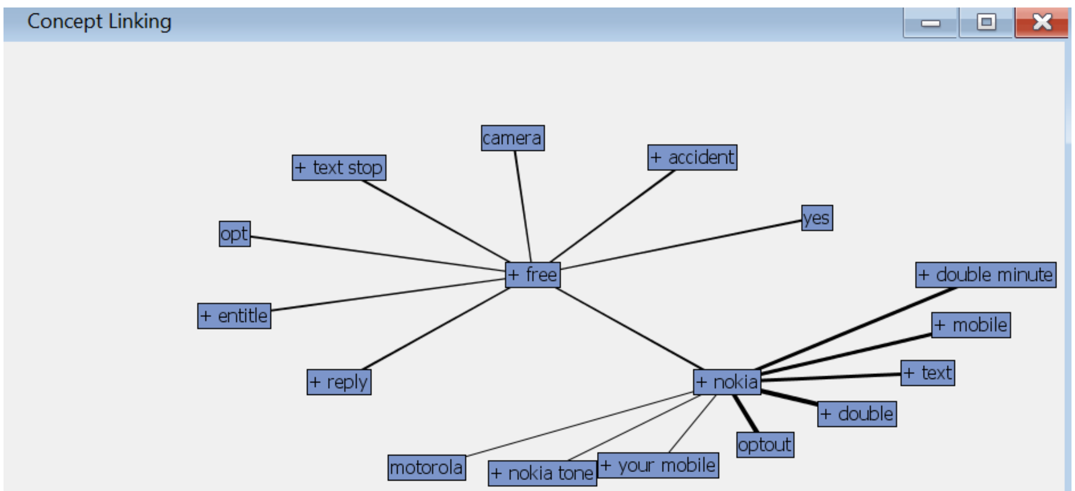


Fig11: Concept link for 'free'

From Fig11, 'Free' is highly associated with 'Nokia', which means spammers are sending messages to customers that they are entitled to get a 'free Nokia' mobile and can 'opt out' from any 'double minute' plan.

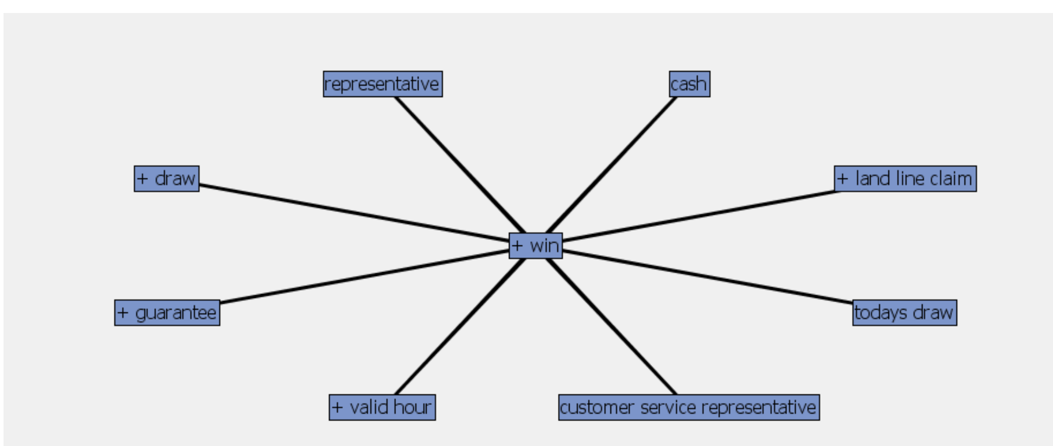


Fig12: Concept link for 'win'

From Fig12, the concept link for the word ‘win’ has a high association with the words ‘cash’, ‘guarantee’ and ‘draw’. This means that spammers send messages to their customers saying they could win a guaranteed cash prize via draws.

Concept link for Ham data

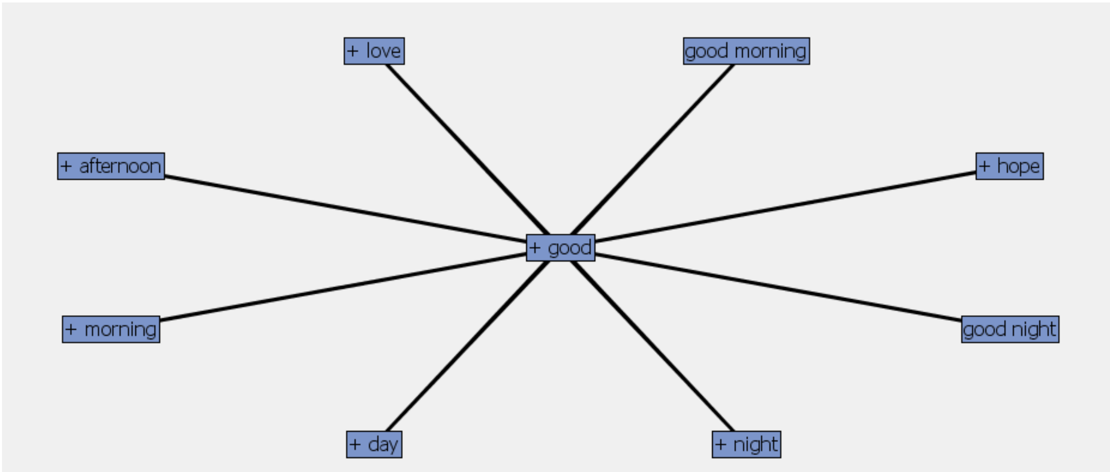


Fig13: Concept link for ‘good’

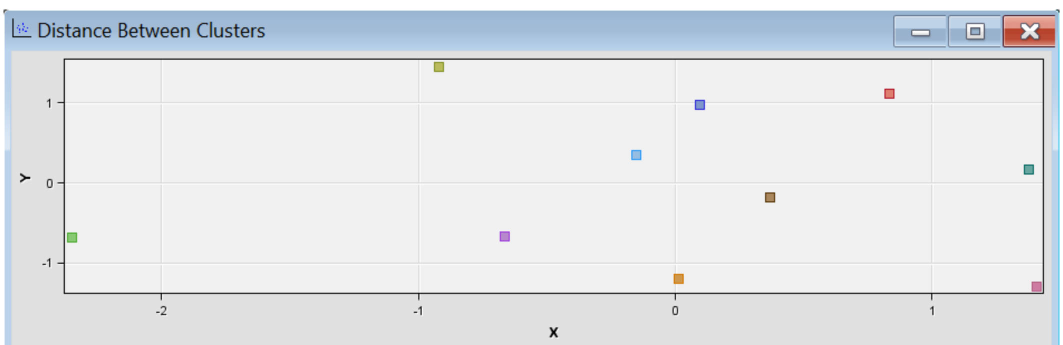
From Fig13, ‘good’ is strongly associated with the words ‘day’, ‘night’, ‘afternoon’ which is not surprising because we generally tend to greet in any regular conversation.

Text Clustering

Once the text has been filtered using the Text Filter node, similar terms in the dataset are grouped together. SAS® Enterprise Miner™ allows to group terms closely related to each other into separate clusters of related terms. After some trial-and-error, the properties settings for the Text Cluster node are set to generate well separated clusters in the cluster space. An exact 10 cluster solution for spam data and 5 cluster solution for ham data using Expectation Maximization Cluster Algorithm and 8 descriptive terms that describe the cluster are generated.

Spam data

The ten clusters generated are well separated from each other and comprise of the terms as seen in Fig14.



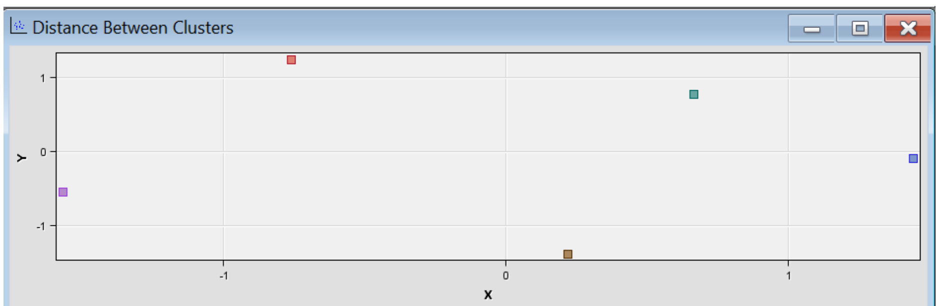
Clusters

Cluster ID	Descriptive Terms
2	+text +free +nokia +phone +mobile +late +tone +month
7	+UK +debt +loan +help +limit http +'world wide web' +info
5	+claim +win +prize +guarantee urgent +contact +'valid hour' +land
6	+accident +entitle +claim +pound +message +reply +record +'text stop'
1	stop +message +sms +dog +love +reply +end +night
4	+week +'terms and conditions' +'world wide web' +voucher promotion entry weekly +chance
3	+service +customer +charge +order +reference ringtone +announcement +arrive
9	urgent +landline +await +holiday cash +'terms and conditions' +award collection
8	+contact 'find out' +'date service' +date +service +'po box' +know +reveal
10	'account statement' account private statement +expire +point +show code

Fig14: Terms describing the spam clusters that are well separated

Ham data

The five clusters generated are well separated from each other as seen in Fig15.



Clusters

Cluster ID	Descriptive Terms
2	+minute yeah +lol +buy +yes +watch +hour +right
3	+want +don't +know +late +message +pick +phone +right
4	+good +day +love +back +today +hope +dont +work
1	+good +know +want +late +don't +love +'let me' +lot
5	+darling +time +tomorrow home +meet +night +leave +reach

Fig15: Terms describing the ham clusters that are well separated

Text Topic

After connecting the Text Filter node in SAS® Enterprise Miner™, the Text Topic node is joined, which enables to combine the terms into topics for further analysis. The properties settings for the Text Topic node have been set to generate same number of topics as the number of clusters generated by the text cluster node for both spam and ham data.

Topic ID	Topic	Number of Terms	# Docs ▼
2	+text,+number,+claim,promotion,+chat	22	399
8	+reply,stop,+minute,+sms,+video	47	302
5	+message,+free,+number,urgent,+waiting	31	282
6	+free,+minute,+phone,+text,+nokia	43	278
10	+number,+service,+customer,cash,+claim	71	249
9	+week,+win,+world wide web,+free,+tone	53	241
1	+claim,+accident,+entitle,+pound,+record	20	197
3	+prize,+win,urgent,+claim,+guarantee	21	196
4	+service,+contact,+date,+date service,+know	27	133
7	+show,code,account,+expire,private	14	49

Fig16: Text topic node results from spam data with custom settings

Topic ID	Topic	Number of Terms	# Docs ▼
5	+love,+day,home,+miss,+darling	38	463
4	+want,+don't,+darling,+time,home	28	382
1	+good,+day,+morning,+night,+hope	17	346
2	+know,+don't,+dont,+let me,+day	12	281
3	+late,+meet,+tomorrow,i am,home	11	224

Fig17: Text topic node results from ham data with custom settings

Rule Based Model

Methodology

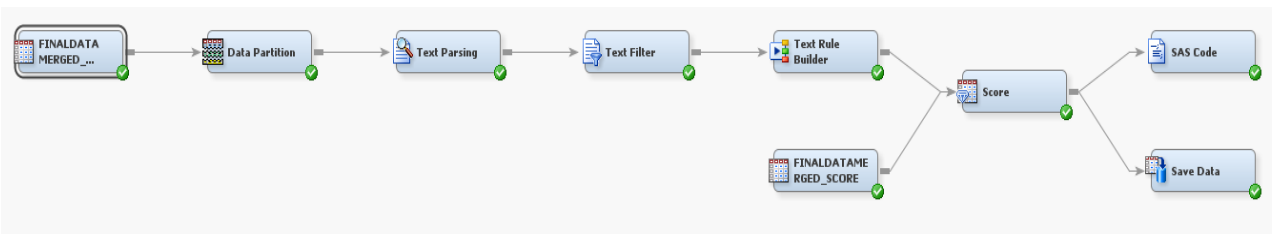


Fig18: Rule based model

The data sets used for this analysis are:

- Finaldatamerged_model.sas7bdat (a dataset that combines spam and ham messages which is 90% of all the messages)
- Finaldatamerged_score.sas7bdat (a dataset that combines spam and ham messages which is 10% of all the messages)
- spam_stopwords_manual.sas7bdat
- engdict.sas7bdat
- spam_syn_manual.sas7bdat
- syn_py1_dropped.sas7bat

Dataset

Since the data is available as a single SAS file, for the purpose of this analysis, the data set has been divided into model data and score data using stratified sampling. Stratified sampling is used to split the data into model and score datasets in the same proportion as the total data. 90% of the total data is considered for model building and 10% of the total data is set aside for scoring. These 2 data sets are added as input sources in SAS Enterprise Miner.

Frequency distribution of total data

The FREQ Procedure

Target				
Target	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ham	4827	69.68	4827	69.68
spam	2100	30.32	6927	100.00

Frequency distribution of model data

The FREQ Procedure

Target				
Target	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ham	4345	69.69	4345	69.69
spam	1890	30.31	6235	100.00

Frequency distribution of scoring data

The FREQ Procedure

Target				
Target	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ham	482	69.65	482	69.65
spam	210	30.35	692	100.00

Fig19: Frequency distributions in total, model, scoring datasets

90% of the stratified sample has 4345 ham messages and 1890 spam messages and the target variable 'spam' and 'ham' is used for the purpose of this analysis. A data partition node is used to set 80% of the observations as training and the rest 20% as validation. Then the text parsing and text filter nodes are added similar to before. All the properties of the text parsing and text filter node are set the same way as before, for building the clusters.

Name	Role	Level
key	ID	Interval
SelectionProb	Rejected	Interval
SamplingWeight	Rejected	Interval
Target	Target	Nominal
Text	Text	Nominal

Fig20: Variable description in Finaldatamerged_model data

Name	Role	Level
Target	Target	Nominal
Text	Text	Nominal
key	ID	Interval

Fig21: Variable description in Finaldatamerged_score data

Text Rule Builder

After the data partition node, text parsing node and the text filter node, next a text rule builder node is added with default combination of settings in the properties panel. The misclassification rate for the validation data is 6%. Text Rule Builder node generates an ordered set of rules that together are useful in describing and predicting a target variable.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Target	Target	ASE	Average Squared Error	0.003255	0.003258	.
Target	Target	DIV	Divisor for ASE	9972	2498	.
Target	Target	MAX	Maximum Absolute Error	0.602614	0.484952	.
Target	Target	NOBS	Sum of Frequencies	4986	1249	.
Target	Target	RASE	Root Average Squared Error	0.057049	0.057075	.
Target	Target	SSE	Sum of Squared Errors	32.45479	8.137272	.
Target	Target	DISF	Frequency of Classified Cases	4986	1249	.
Target	Target	MISC	Misclassification Rate	0.033293	0.060849	.
Target	Target	WRONG	Number of Wrong Classifications	166	76	.

Fig22: Fit statistics for the text rule builder model

Rules Obtained

Target Value	Rule #	Rule	Precision
HAM	1	sleep	100.0%
HAM	2	lol	100.0%
HAM	3	morning	100.0%
HAM	4	watch	100.0%
HAM	5	finish	100.0%
HAM	6	alright	100.0%
HAM	7	yeah	100.0%
HAM	8	don't & ~chat & ~minute	99.80%
HAM	9	eat	99.81%
HAM	10	gonna	99.82%
HAM	11	darling	99.40%
HAM	12	happen	99.43%
HAM	13	yup	99.45%
HAM	14	love & ~text & ~chat	99.19%
HAM	15	leave & ~message	99.13%

Rules Obtained

Target Value	Rule #	Rule	Precision
SPAM	66	claim	100.0%
SPAM	67	service	99.60%
SPAM	68	world wide web	99.36%
SPAM	69	text & reply	99.42%
SPAM	70	your mobile	99.46%
SPAM	71	tone	99.49%
SPAM	72	terms and conditions	99.28%
SPAM	73	promotion	99.31%
SPAM	74	optout	99.33%
SPAM	75	UK	99.35%
SPAM	76	prize	99.37%
SPAM	77	account statement	99.38%
SPAM	78	immediately	99.39%
SPAM	79	po box	99.41%
SPAM	80	landline	99.32%
SPAM	81	text & free	99.25%
SPAM	82	debt	99.26%

Fig23: Rules to classify spam and ham messages

The most important rule (rule # 66) is, if the message contains the term 'claim' then the message can be classified as spam and if the message contains 'don't' without 'chat' or 'minute' then it can be classified as ham (rule # 8).

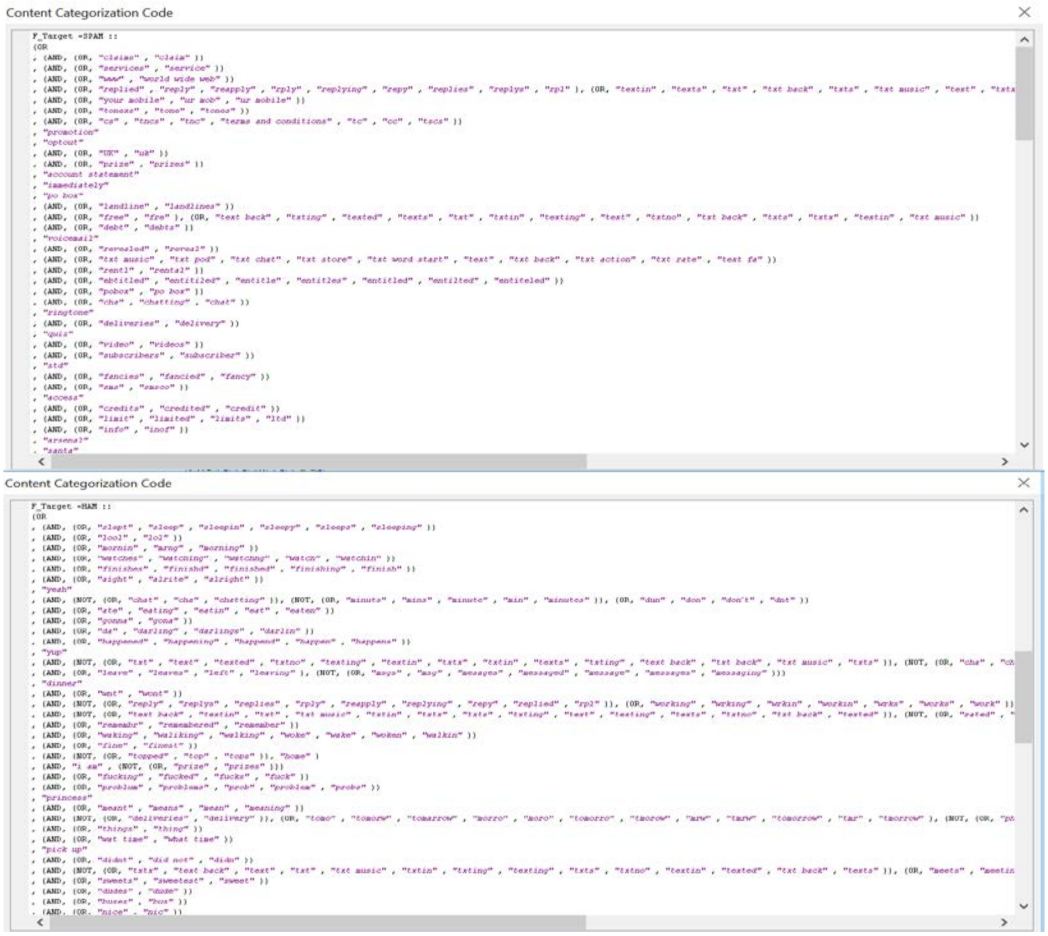


Fig24: Content categorization code obtained from the text rule-builder node

While the model seems to be performing reasonably good from looking at the overall misclassification rate which is 6%, the model classifies each outcome (spam or ham) reasonably well in both spam and ham datasets. The numbers reported below show that the model does about equally well in predicting positive versus negative cases.

Output

Classification Table

Data Role=TRAIN Target Variable=Target Target Label=Target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
HAM	HAM	96.1249	99.2230	3448	69.1536
SPAM	HAM	3.8751	9.1992	139	2.7878
HAM	SPAM	1.9299	0.7770	27	0.5415
SPAM	SPAM	98.0701	90.8008	1372	27.5170

Data Role=VALIDATE Target Variable=Target Target Label=Target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
HAM	HAM	93.5307	98.0460	853	68.2946
SPAM	HAM	6.4693	15.5673	59	4.7238
HAM	SPAM	5.0445	1.9540	17	1.3611
SPAM	SPAM	94.9555	84.4327	320	25.6205

Fig25: Model Classification Results from the Rule-Builder Node for Spam and Ham messages

Scoring

Now using the model build, the data set aside to score is scored. There are a total of 482 ham and 210 spam messages in the scoring data.

The scoring results shown below look reasonable, since the % of spam and ham in the scored data is similar to those from the training and validation data. However, in this scored data set (unlike in real scoring cases), we have the actual target (spam or ham) values, and those can be compared against the predicted target from the text rule-builder model via a cross-tab. The cross-tab between the two results can be generated easily by using a SAS code node in this diagram space.

```
Class Variable Summary Statistics

Data Role=SCORE Output Type=CLASSIFICATION

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent

I_Target       .          HAM          496        71.6763
I_Target       .          SPAM         196        28.3237


Data Role=TRAIN Output Type=CLASSIFICATION

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent

I_Target       .          HAM          3587       71.9414
I_Target       .          SPAM         1399       28.0586


Data Role=VALIDATE Output Type=CLASSIFICATION

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent

I_Target       .          HAM          912        73.0184
I_Target       .          SPAM         337        26.9816
```

Fig26: Scoring results

SAS code node

The scored data set which now has both actual target variable and predicted variable can be used to perform cross tabs to get a sense of how many actual targets are present and how many of them are correctly being predicted.

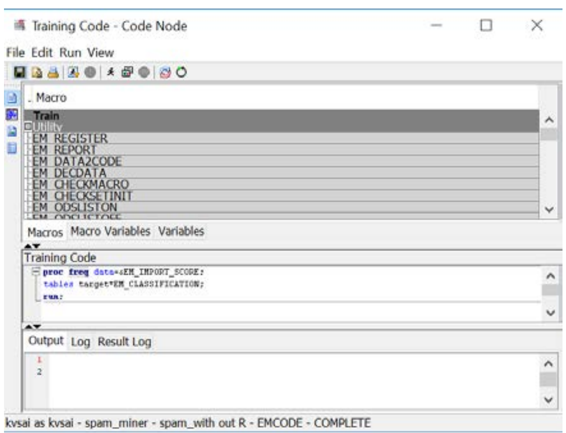


Fig27: SAS code for finding cross tabs between target (original target) and EM_CLASSIFICATION (Prediction for target)

It seems that 475 out of 482 ham messages (98.55%) were correctly classified, and 189 out of 210 spam messages (90%) were also correctly classified. Overall, 664 out of 692 (95.95%) messages were correctly classified by the text rule builder model.

```
The FREQ Procedure

Table of Target by EM_CLASSIFICATION

Target(Target)
              EM_CLASSIFICATION(Prediction for Target)

Frequency    |
Percent      |
Row Pct      |
Col Pct      |HAM     |SPAM    |  Total
-------------+--------+--------+
ham          |    475 |      7 |    482
             |  68.64 |   1.01 |  69.65
             |  98.55 |   1.45 |
             |  95.77 |   3.57 |
-------------+--------+--------+
spam         |     21 |    189 |    210
             |   3.03 |  27.31 |  30.35
             |  10.00 |  90.00 |
             |   4.23 |  96.43 |
-------------+--------+--------+
Total             496      196      692
                71.68    28.32   100.00
```

Fig28: Comparing scoring results with known values.

Conclusion

Identifying if a message is either ham or spam, can be very helpful to both customers and companies. Carrier companies can protect their customers from spammers and their spam texts. Companies can use the list of high frequency spam words and take necessary precautions to not include these words in their promotional offers. Score node can be used to test new messages. They can be predicted as spam or ham with the help of text rule builder node.

From the concept link for text, we observe that spammers are asking their customers to text back if they want either free minutes or a camcorder or a video phone. From the concept link for free, we observe that spammers are sending messages to customers that they are entitled to get a free Nokia mobile phones and can opt out from any double minute plan. From the concept link for the word win, which has a high association with the words cash, guarantee and draw, spammers send messages to their customers saying they could win a guaranteed cash prize via draws.

From the concept link for good for ham message, which is strongly associated with the words day, night, afternoon, because we generally tend to greet in any regular conversation. Using text builder rules, if a message contains the term ‘claim’ then the message can be classified as spam and if the message contains ‘don’t’ without ‘chat’ or ‘minute’ then it can be classified as ham.

Limitations and Future work

All the data has not occurred in the same linguistic region. While the spam data is in British English and is drawn from 2 UK public consumer complaints websites, the non-spam is a combination of data from two very disparate sources. The NUS non spam data is strongly influenced by Singaporean English.

The distribution of spam and non-spam in the corpus is arbitrary and the actual distribution of spam can only be found by analyzing a full stream of SMS traffic.

References

Eric A. Taub. "Fighting Back Against Spam Texts" *The New York Times*, 4 April 2012, www.nytimes.com/2012/04/05/technology/personaltech/fighting-back-against-spam-texts.html?mcubz=3. Accessed 7 Feb 2017.

Sarah Jane Delany, Mark Buckley, Derek Greene, "SMS Spam Filtering: Methods and Data" *Dublin Institute of Technology*, Feb 2012, arrow.dit.ie/scschcomart/17/. Accessed 7 Feb 2017.

UCI Machine learning, *Kaggle SMS Spam Collection Dataset: Collection of SMS Messages Tagged as Spam or Legitimate*, Jan 2017, www.kaggle.com/uciml/sms-spam-collection-dataset. Accessed Jan 27 2017.

Tiago A. Almeida, Jose Maria Gomez Hidalgo, "SMS Spam Collection v. 1", 2013, www.dt.fee.unicamp.br/~tiago/smsspamcollection/. Accessed 7 Feb 2017.

Goutam Chakraborty, Murali Pagolu, Satish Garla, *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® by SAS Institute Inc 2014*, Accessed 20 Mar 2017.

Getting Started with SAS® Text Miner 13.2. Cary, NC: SAS Institute Inc., Accessed 20 Mar 2017.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Mounika Kondamudi
Oklahoma State University
Phone: (405) 564 3633
Email: mounika.kondamudi@okstate.edu
LinkedIn: https://www.linkedin.com/in/mounikakondamudi

Mounika Kondamudi is a Graduate student in Business Analytics at Oklahoma State University where she is currently working as a Graduate Research Assistant in Marketing Department. She has worked as Product Research and Analysis Intern on Credit and Finance projects with CreditXpert, Maryland. She is a SAS® Certified Base programmer and a SAS® Certified Advance programmer. She is a scholarship winner for her poster at Analytics Experience, 2017 at Washington DC and a scholarship winner for her paper presentation at the SCSUG conference, 2017 at Dallas.