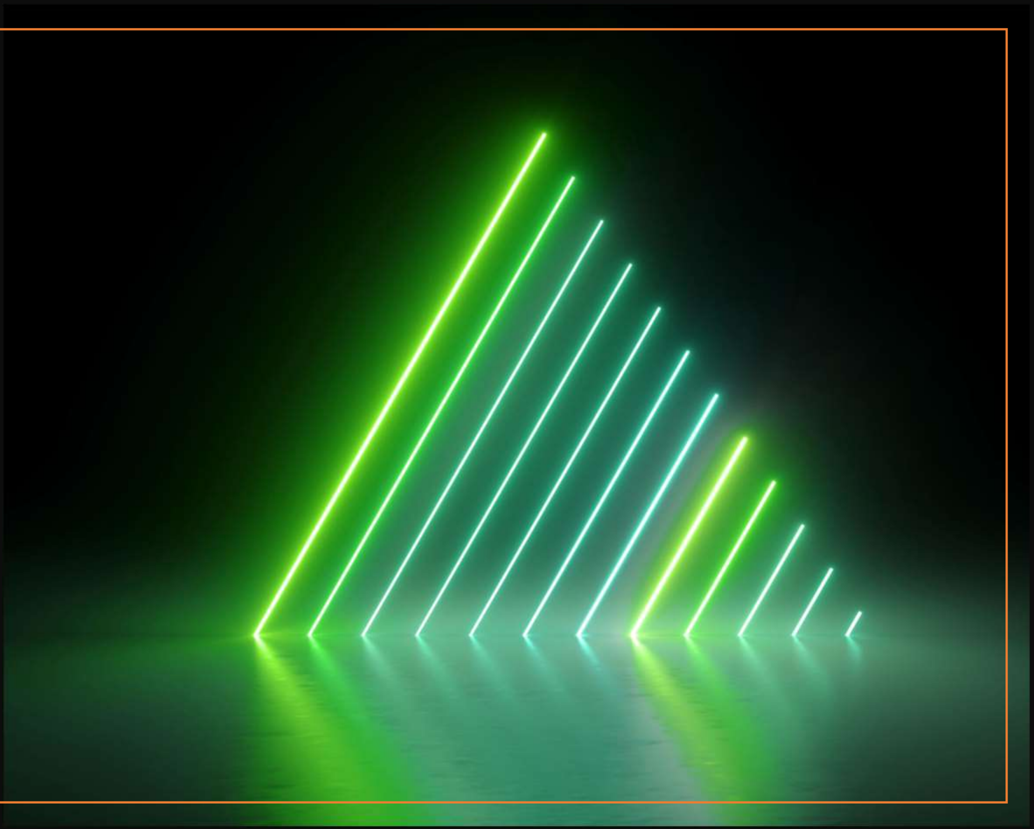
Text Analytics

Topic: Text Mining and Feature Selection

Sowmiya Muruganandam,



Text Mining

Case in Point: Predict customer churn using Text mining

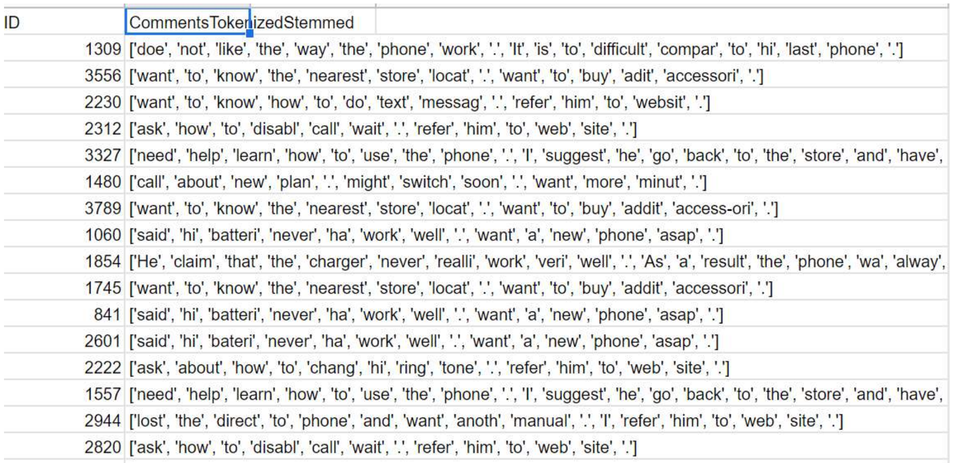
The process to convert unstructured data to normalized structured data using Natural Language Processing

Problem Approach:

1) **Tokenize the data**
2) **Stem tokens/words**
3) **Stemmer Comparison**
4) **Implementation Bag of words model**
5) **Combine text and customer data (semi-structured to structured conversion complete)**
6) **One hot encode the categorical variables**
7) **Use feature selection to determine best features**
8) **Split data (80%-20%) and build classification models**

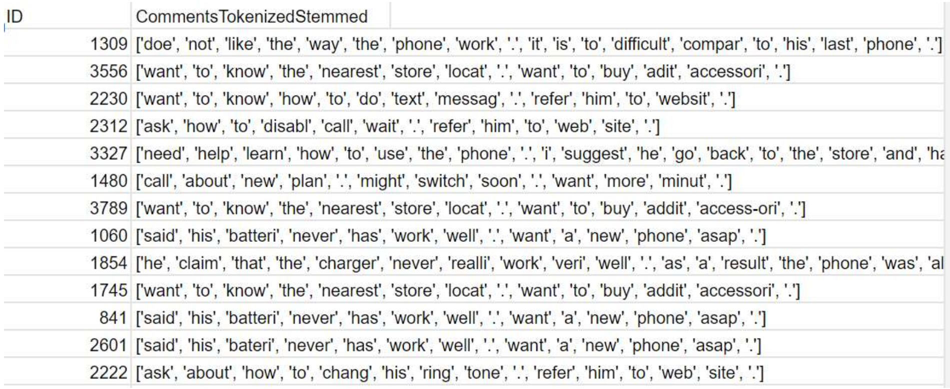
Stemmer Comparison

Snowball



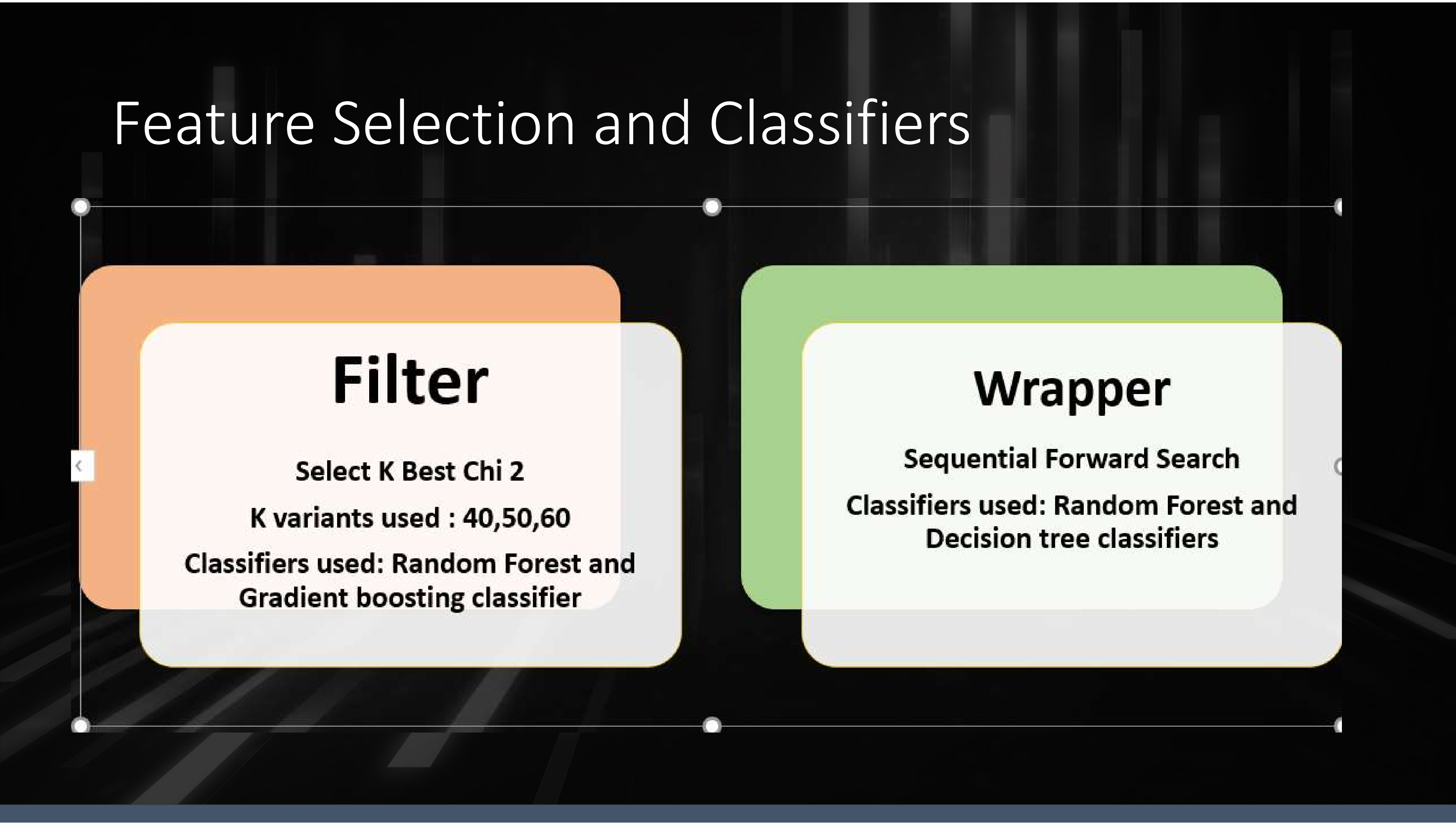
ID	CommentsTokenizedStemmed
1309	['doe', 'not', 'like', 'the', 'way', 'the', 'phone', 'work', '.', 'It', 'is', 'to', 'difficult', 'compar', 'to', 'hi', 'last', 'phone', '.']
3556	['want', 'to', 'know', 'the', 'nearest', 'store', 'locat', '.', 'want', 'to', 'buy', 'adit', 'accessori', '.']
2230	['want', 'to', 'know', 'how', 'to', 'do', 'text', 'messag', '.', 'refer', 'him', 'to', 'websit', '.']
2312	['ask', 'how', 'to', 'disabl', 'call', 'wait', '.', 'refer', 'him', 'to', 'web', 'site', '.']
3327	['need', 'help', 'learn', 'how', 'to', 'use', 'the', 'phone', '.', 'I', 'suggest', 'he', 'go', 'back', 'to', 'the', 'store', 'and', 'have',
1480	['call', 'about', 'new', 'plan', '.', 'might', 'switch', 'soon', '.', 'want', 'more', 'minut', '.']
3789	['want', 'to', 'know', 'the', 'nearest', 'store', 'locat', '.', 'want', 'to', 'buy', 'addit', 'access-ori', '.']
1060	['said', 'hi', 'batteri', 'never', 'ha', 'work', 'well', '.', 'want', 'a', 'new', 'phone', 'asap', '.']
1854	['He', 'claim', 'that', 'the', 'charger', 'never', 'realli', 'work', 'veri', 'well', '.', 'As', 'a', 'result', 'the', 'phone', 'wa', 'alway',
1745	['want', 'to', 'know', 'the', 'nearest', 'store', 'locat', '.', 'want', 'to', 'buy', 'addit', 'accessori', '.']
841	['said', 'hi', 'batteri', 'never', 'ha', 'work', 'well', '.', 'want', 'a', 'new', 'phone', 'asap', '.']
2601	['said', 'hi', 'bateri', 'never', 'ha', 'work', 'well', '.', 'want', 'a', 'new', 'phone', 'asap', '.']
2222	['ask', 'about', 'how', 'to', 'chang', 'hi', 'ring', 'tone', '.', 'refer', 'him', 'to', 'web', 'site', '.']
1557	['need', 'help', 'learn', 'how', 'to', 'use', 'the', 'phone', '.', 'I', 'suggest', 'he', 'go', 'back', 'to', 'the', 'store', 'and', 'have',
2944	['lost', 'the', 'direct', 'to', 'phone', 'and', 'want', 'anoth', 'manual', '.', 'I', 'refer', 'him', 'to', 'web', 'site', '.']
2820	['ask', 'how', 'to', 'disabl', 'call', 'wait', '.', 'refer', 'him', 'to', 'web', 'site', '.']

Porter



ID	CommentsTokenizedStemmed
1309	['doe', 'not', 'like', 'the', 'way', 'the', 'phone', 'work', '.', 'it', 'is', 'to', 'difficult', 'compar', 'to', 'his', 'last', 'phone', '.']
3556	['want', 'to', 'know', 'the', 'nearest', 'store', 'locat', '.', 'want', 'to', 'buy', 'adit', 'accessori', '.']
2230	['want', 'to', 'know', 'how', 'to', 'do', 'text', 'messag', '.', 'refer', 'him', 'to', 'websit', '.']
2312	['ask', 'how', 'to', 'disabl', 'call', 'wait', '.', 'refer', 'him', 'to', 'web', 'site', '.']
3327	['need', 'help', 'learn', 'how', 'to', 'use', 'the', 'phone', '.', 'i', 'suggest', 'he', 'go', 'back', 'to', 'the', 'store', 'and', 'ha
1480	['call', 'about', 'new', 'plan', '.', 'might', 'switch', 'soon', '.', 'want', 'more', 'minut', '.']
3789	['want', 'to', 'know', 'the', 'nearest', 'store', 'locat', '.', 'want', 'to', 'buy', 'addit', 'access-ori', '.']
1060	['said', 'his', 'batteri', 'never', 'has', 'work', 'well', '.', 'want', 'a', 'new', 'phone', 'asap', '.']
1854	['he', 'claim', 'that', 'the', 'charger', 'never', 'realli', 'work', 'veri', 'well', '.', 'as', 'a', 'result', 'the', 'phone', 'was', 'al
1745	['want', 'to', 'know', 'the', 'nearest', 'store', 'locat', '.', 'want', 'to', 'buy', 'addit', 'accessori', '.']
841	['said', 'his', 'batteri', 'never', 'has', 'work', 'well', '.', 'want', 'a', 'new', 'phone', 'asap', '.']
2601	['said', 'his', 'bateri', 'never', 'has', 'work', 'well', '.', 'want', 'a', 'new', 'phone', 'asap', '.']
2222	['ask', 'about', 'how', 'to', 'chang', 'his', 'ring', 'tone', '.', 'refer', 'him', 'to', 'web', 'site', '.']

Snowball is better than Porter



Feature Selection and Classifiers

Filter

Select K Best Chi 2

K variants used : 40,50,60

Classifiers used: Random Forest and Gradient boosting classifier

Wrapper

Sequential Forward Search

Classifiers used: Random Forest and Decision tree classifiers

Accuracy Scores Obtained

Feature Selection	Classifier	Accuracy Score
Filter: Chi Squared K=40	Random Forest	0.86714
Filter: Chi Squared K=40	Gradient Boosting	0.874396
Filter: Chi Squared K=50	Random Forest	0.86231
Filter: Chi Squared K=50	Gradient Boosting	0.8405
Filter: Chi Squared K=60	Random Forest	0.87922
Filter: Chi Squared K=60	Gradient Boosting	0.85507
Wrapper : Sequential Forward Search	Random Forest	0.57971
Wrapper : Sequential Forward Search	Decision Tree	0.5700

Thank You!