sms-spam-v1-stop

April 23, 2019

V1: + Delete the stop words + All models uses the TfidfVectorizer to do the the preprocessing Goal of this notebook to test several classifiers on the data set with different features

0.0.1 Let's begin

First of all neccesary imports

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib as mpl
        import matplotlib.pyplot as plt
        import seaborn as sns
        import string
        from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.model_selection import train_test_split
        from nltk.corpus import stopwords
        from sklearn.preprocessing import LabelEncoder
        %matplotlib inline
```

Let's read the data from csv file

```
In [2]: sms = pd.read_csv('../data/sms-spam.csv',delimiter=',',encoding='latin-1')

        sms.head()
```

```
Out[2]:      v1                                                 v2 Unnamed: 2  \
        0   ham  Go until jurong point, crazy.. Available only ...        NaN
        1   ham                      Ok lar... Joking wif u oni...        NaN
        2  spam  Free entry in 2 a wkly comp to win FA Cup fina...        NaN
        3   ham  U dun say so early hor... U c already then say...        NaN
        4   ham  Nah I don't think he goes to usf, he lives aro...        NaN

          Unnamed: 3 Unnamed: 4
        0        NaN        NaN
        1        NaN        NaN
        2        NaN        NaN
        3        NaN        NaN
        4        NaN        NaN
```

Now drop "unnamed" columns and rename v1 and v2 to "label" and "message"

```
In [3]: sms = sms.drop(['Unnamed: 2','Unnamed: 3','Unnamed: 4'],axis=1)
        sms = sms.rename(columns = {'v1':'label','v2':'message'})
```

Let's look into our data

```
In [4]: sms.groupby('label').describe()
```

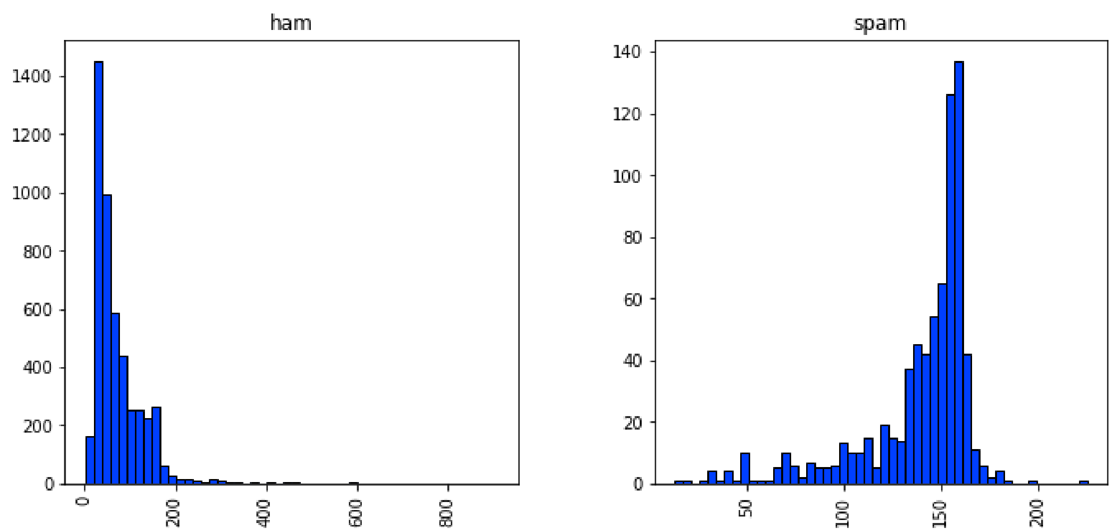
```
Out[4]:       message
               count unique                                              top freq
       label
       ham      4825   4516                            Sorry, I'll call later   30
       spam      747    653  Please call our customer service representativ...    4
```

Intresting that "Sorry, I'll call later" appears only 30 times here =)
Now let's create new feature "message length" and plot it to see if it's of any interest

```
In [5]: sms['length'] = sms['message'].apply(len)
        sms.head()
```

```
Out[5]:   label                                            message  length
       0   ham  Go until jurong point, crazy.. Available only ...     111
       1   ham                      Ok lar... Joking wif u oni...      29
       2  spam  Free entry in 2 a wkly comp to win FA Cup fina...     155
       3   ham  U dun say so early hor... U c already then say...      49
       4   ham  Nah I don't think he goes to usf, he lives aro...      61
```

```
In [6]: mpl.rcParams['patch.force_edgecolor'] = True
        plt.style.use('seaborn-bright')
        sms.hist(column='length', by='label', bins=50,figsize=(11,5))
        plt.savefig("../img/sms-length-distribution.eps")
        plt.show()
```



Looks like the lengthy is the message, more likely it is a spam. Let's not forget this

0.0.2 Text processing and vectorizing our meddages

Let's create new data frame. We'll need a copy later on

```
In [7]: text_feat = sms['message'].copy()
```

Now define our tex precessing function. It will remove any punctuation and stopwords aswell.

```
In [8]: def text_process(text):

            text = text.translate(str.maketrans('', '', string.punctuation))
            text = [word for word in text.split() if word.lower() not in stopwords.words('engli

            return " ".join(text)
```

```
In [9]: text_feat = text_feat.apply(text_process)
```

```
In [10]: vectorizer = TfidfVectorizer("english")
```

```
In [11]: features = vectorizer.fit_transform(text_feat)
```

```
In [12]: labels = LabelEncoder().fit_transform(sms['label'])
         labels = labels.reshape(-1,1)
```

```
In [13]: text_feat.shape
```

```
Out[13]: (5572,)
```

```
In [14]: features.shape
```

```
Out[14]: (5572, 9403)
```

0.0.3 Classifiers and predictions

First of all let's split our features to test and train set

```
In [15]: features_train, features_test, labels_train, labels_test = train_test_split(features,
```

Now let's import bunch of classifiers, initialize them and make a dictionary to itereate through

```
In [16]: from sklearn.linear_model import LogisticRegression
         from sklearn.svm import SVC
         from sklearn.naive_bayes import MultinomialNB
         from sklearn.tree import DecisionTreeClassifier
         from sklearn.neighbors import KNeighborsClassifier
         from sklearn.ensemble import RandomForestClassifier
         from sklearn.ensemble import AdaBoostClassifier
         from sklearn.ensemble import BaggingClassifier
         from sklearn.ensemble import ExtraTreesClassifier
         from sklearn.metrics import precision_score, recall_score, accuracy_score, f1_score
```

```
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/ensemble/weight_boosting.py:29
  from numpy.core.umath_tests import inner1d
```

```
In [17]: svc = SVC(kernel='sigmoid', gamma=1.0)
         knc = KNeighborsClassifier(n_neighbors=49)
         mnb = MultinomialNB(alpha=0.2)
         dtc = DecisionTreeClassifier(min_samples_split=7, random_state=111)
         lrc = LogisticRegression(solver='liblinear', penalty='l1')
         rfc = RandomForestClassifier(n_estimators=31, random_state=111)
         abc = AdaBoostClassifier(n_estimators=62, random_state=111)
         bc = BaggingClassifier(n_estimators=9, random_state=111)
         etc = ExtraTreesClassifier(n_estimators=9, random_state=111)
```

```
In [18]: clfs = {'SVC' : svc,'KN' : knc, 'NB': mnb, 'DT': dtc, 'LR': lrc, 'RF': rfc, 'AdaBoost
```

Let's make functions to fit our classifiers and make predictions

```
In [19]: def train_classifier(clf, feature_train, labels_train):
             clf.fit(feature_train, labels_train)
```

```
In [20]: def predict_labels(clf, features):
             return (clf.predict(features))
```

Now iterate through classifiers and save the results

```
In [21]: import time
```

```
In [22]: pred_scores = []
         for k,v in clfs.items():
             since = time.time()

             train_classifier(v, features_train, labels_train)
             time_elapsed = time.time() - since

             pred = predict_labels(v,features_test)
             pred_scores.append((k, [precision_score(labels_test,pred), recall_score(labels_tes
```

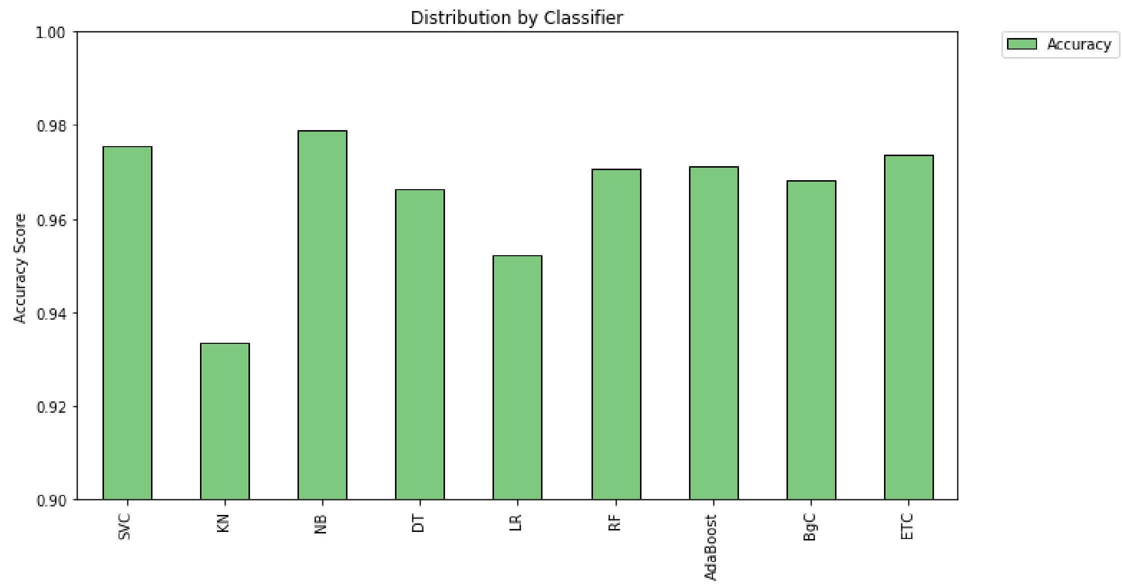
```
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/utils/validation.py:578: DataC
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/ipykernel/__main__.py:2: DataConversio
  from ipykernel import kernelapp as app
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/ipykernel/__main__.py:2: DataConversio
  from ipykernel import kernelapp as app
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/utils/validation.py:578: DataC
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/utils/validation.py:578: DataC
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/ipykernel/__main__.py:2: DataConversio
  from ipykernel import kernelapp as app
```

```
In [23]: # pred_scores
```

```
In [24]: df = pd.DataFrame.from_items(pred_scores,orient='index', columns=['Precision', 'Recal
         df
```

```
Out[24]:            Precision    Recall  Accuracy        F1 Training Time (s)
         SVC         0.989305  0.825893  0.975478  0.900243         0m 0.3595s
         KN          0.991304  0.508929  0.933612  0.672566         0m 0.0031s
         NB          0.939535  0.901786  0.979067  0.920273         0m 0.0019s
         DT          0.878378  0.870536  0.966507  0.874439         0m 0.2016s
         LR          0.967532  0.665179  0.952153  0.788360         0m 0.0097s
         RF          1.000000  0.781250  0.970694  0.877193         0m 1.2949s
         AdaBoost    0.948980  0.830357  0.971292  0.885714         0m 2.5555s
         BgC         0.905213  0.852679  0.968301  0.878161         0m 1.0332s
         ETC         0.978723  0.821429  0.973684  0.893204         0m 0.9084s
```

```
In [25]: df.plot(kind='bar', y="Accuracy", ylim=(0.9,1.0), figsize=(11,6), align='center', colo
         plt.xticks(np.arange(9), df.index)
         plt.ylabel('Accuracy Score')
         plt.title('Distribution by Classifier')
         plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
         plt.savefig("../img/sms-acc-basemodel-v1-stop.eps")
         plt.show()
```



Looks like ensemble classifiers are not doing as good as expected.

0.0.4 Voting classifier

We are using ensemble algorithms here, but what about ensemble of ensembles? Will it beat NB?

```
In [26]: from sklearn.ensemble import VotingClassifier

In [27]: eclf = VotingClassifier(estimators=[('BgC', bc), ('ETC', etc), ('RF', rfc), ('Ada', a

In [28]: eclf.fit(features_train,labels_train)
```

```
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/preprocessing/label.py:95: Da
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/preprocessing/label.py:128: D
  y = column_or_1d(y, warn=True)
```

```
Out[28]: VotingClassifier(estimators=[('BgC', BaggingClassifier(base_estimator=None, bootstrap
                  bootstrap_features=False, max_features=1.0, max_samples=1.0,
                  n_estimators=9, n_jobs=1, oob_score=False, random_state=111,
                  verbose=0, warm_start=False)), ('ETC', ExtraTreesClassifier(bootstrap=False,
                   learning_rate=1.0, n_estimators=62, random_state=111))],
                  flatten_transform=None, n_jobs=1, voting='soft', weights=None)
```

```
In [29]: pred = eclf.predict(features_test)
```

```
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: D
  if diff:
```

```
In [30]: print(precision_score(labels_test,pred), recall_score(labels_test,pred), accuracy_sco
```

```
1.0 0.84375 0.979066985645933 0.9152542372881356
```

Better but nope.

0.0.5 RNN

Define the RNN structure.

```
In [31]: from keras.models import Model
         from keras.layers import LSTM, Activation, Dense, Dropout, Input, Embedding
         from keras.optimizers import RMSprop
         from keras.preprocessing.text import Tokenizer
         from keras.preprocessing import sequence
         from keras.utils import to_categorical
         from keras.callbacks import EarlyStopping
         from keras.callbacks import Callback
```

```
Using TensorFlow backend.
```

```
In [32]: max_words = features_train.shape[0]
         max_len = features_train.shape[1]
```

```
In [33]: def RNN():
             inputs = Input(name='inputs',shape=[max_len])
             layer = Embedding(max_words,50,input_length=max_len)(inputs)
             layer = LSTM(64)(layer)
             layer = Dense(256,name='FC1')(layer)
             layer = Activation('relu')(layer)
             layer = Dropout(0.5)(layer)
             layer = Dense(1,name='out_layer')(layer)
             layer = Activation('sigmoid')(layer)
             model = Model(inputs=inputs,outputs=layer)
             return model
```

Call the function and compile the model.

```
In [34]: model = RNN()
         model.summary()
         model.compile(loss='binary_crossentropy',optimizer=RMSprop(),metrics=['accuracy'])
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inputs (InputLayer)          (None, 9403)              0
_________________________________________________________________
embedding_1 (Embedding)      (None, 9403, 50)          195000
_________________________________________________________________
lstm_1 (LSTM)                (None, 64)                29440
_________________________________________________________________
FC1 (Dense)                  (None, 256)               16640
_________________________________________________________________
activation_1 (Activation)    (None, 256)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
out_layer (Dense)            (None, 1)                 257
_________________________________________________________________
activation_2 (Activation)    (None, 1)                 0
=================================================================
Total params: 241,337
Trainable params: 241,337
Non-trainable params: 0
_________________________________________________________________
```

```
In [35]: since = time.time()

         # model.fit(features_train, labels_train, epochs=10, batch_size=128,
         #                        validation_split=0.2,
         #                        callbacks=[metrics, EarlyStopping(monitor='val_loss',min_delt
```

```
        model.fit(features_train,labels_train,batch_size=128,epochs=10,
                  validation_split=0.2,callbacks=[EarlyStopping(monitor='val_loss',min_delta=

        time_elapsed = time.time() - since
```

```
Train on 3120 samples, validate on 780 samples
Epoch 1/10
3120/3120 [==============================] - 298s 96ms/step - loss: 0.4477 - acc: 0.8471 - val
Epoch 2/10
3120/3120 [==============================] - 290s 93ms/step - loss: 0.4053 - acc: 0.8654 - val
```

```
In [36]: print('Training complete in {:.0f}m {:.4f}s'.format(
             time_elapsed // 60, time_elapsed % 60))
```

```
Training complete in 9m 49.5561s
```

```
In [38]: pred = (np.asarray(model.predict(features_test, batch_size=128))).round()

In [42]: pred_scores.append(("LSTM", [precision_score(labels_test,pred), recall_score(labels_te
```

```
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/metrics/classification.py:113
  'precision', 'predicted', average, warn_for)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/metrics/classification.py:113
  'precision', 'predicted', average, warn_for)
```

0.0.6 gcForest

```
In [43]: import sys
         sys.path.append("..")
         from gcforest.gcforest import GCForest
         from gcforest.utils.config_utils import load_json

In [44]: def get_toy_config():
             config = {}
             ca_config = {}
             ca_config["random_state"] = 111
             ca_config["max_layers"] = 10
             ca_config["early_stopping_rounds"] = 3
             ca_config["n_classes"] = 2
             ca_config["estimators"] = []
             ca_config["estimators"].append({"n_folds": 5, "type": "RandomForestClassifier", "
             ca_config["estimators"].append({"n_folds": 5, "type": "RandomForestClassifier", "
             ca_config["estimators"].append({"n_folds": 5, "type": "MultinomialNB", "alpha": 0
             ca_config["estimators"].append({"n_folds": 5, "type": "MultinomialNB", "alpha": 0

             config["cascade"] = ca_config
             return config
```

In [45]:
```
config = get_toy_config()
gc = GCForest(config)

# features_train  ndarraylabels_train  (n_samples, )(n_samples, 1)
features_train = features_train.toarray()
labels_train = labels_train.reshape(-1)

since = time.time()
gc.fit_transform(features_train, labels_train)

time_elapsed = time.time() - since
```

```
[ 2019-04-23 21:00:02,205][cascade_classifier.fit_transform] X_groups_train.shape=[(3900, 9403)
[ 2019-04-23 21:00:02,455][cascade_classifier.fit_transform] group_dims=[9403]
[ 2019-04-23 21:00:02,456][cascade_classifier.fit_transform] group_starts=[0]
[ 2019-04-23 21:00:02,456][cascade_classifier.fit_transform] group_ends=[9403]
[ 2019-04-23 21:00:02,457][cascade_classifier.fit_transform] X_train.shape=(3900, 9403),X_test
[ 2019-04-23 21:00:02,645][cascade_classifier.fit_transform] [layer=0] look_indexs=[0], X_cur_
[ 2019-04-23 21:00:06,792][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 21:00:10,442][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 21:00:13,723][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 21:00:17,217][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 21:00:20,761][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 21:00:20,762][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 21:00:24,097][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 21:00:27,550][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 21:00:30,885][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 21:00:34,334][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 21:00:37,672][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 21:00:37,673][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 21:00:37,909][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_2 - 5_
[ 2019-04-23 21:00:38,141][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_2 - 5_
[ 2019-04-23 21:00:38,367][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_2 - 5_
[ 2019-04-23 21:00:38,593][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_2 - 5_
[ 2019-04-23 21:00:38,822][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_2 - 5_
[ 2019-04-23 21:00:38,823][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_2 - 5_
[ 2019-04-23 21:00:39,049][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_3 - 5_
[ 2019-04-23 21:00:39,276][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_3 - 5_
[ 2019-04-23 21:00:39,501][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_3 - 5_
[ 2019-04-23 21:00:39,722][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_3 - 5_
[ 2019-04-23 21:00:39,946][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_3 - 5_
[ 2019-04-23 21:00:39,947][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_3 - 5_
[ 2019-04-23 21:00:39,948][cascade_classifier.calc_accuracy] Accuracy(layer_0 - train.classifi
[ 2019-04-23 21:00:40,154][cascade_classifier.fit_transform] [layer=1] look_indexs=[0], X_cur_
[ 2019-04-23 21:00:42,324][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 21:00:44,332][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 21:00:46,486][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 21:00:48,549][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
```

```
[ 2019-04-23 21:00:50,349][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 21:00:50,350][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 21:00:52,456][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 21:00:54,561][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 21:00:56,455][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 21:00:58,472][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 21:01:00,569][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 21:01:00,571][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 21:01:00,810][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_2 - 5_
[ 2019-04-23 21:01:01,038][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_2 - 5_
[ 2019-04-23 21:01:01,267][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_2 - 5_
[ 2019-04-23 21:01:01,491][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_2 - 5_
[ 2019-04-23 21:01:01,722][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_2 - 5_
[ 2019-04-23 21:01:01,723][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_2 - 5_
[ 2019-04-23 21:01:01,965][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_3 - 5_
[ 2019-04-23 21:01:02,191][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_3 - 5_
[ 2019-04-23 21:01:02,419][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_3 - 5_
[ 2019-04-23 21:01:02,644][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_3 - 5_
[ 2019-04-23 21:01:02,876][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_3 - 5_
[ 2019-04-23 21:01:02,877][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_3 - 5_
[ 2019-04-23 21:01:02,878][cascade_classifier.calc_accuracy] Accuracy(layer_1 - train.classifi
[ 2019-04-23 21:01:03,059][cascade_classifier.fit_transform] [layer=2] look_indexs=[0], X_cur_
[ 2019-04-23 21:01:04,857][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 21:01:06,762][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 21:01:08,654][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 21:01:10,437][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 21:01:12,226][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 21:01:12,227][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 21:01:14,023][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 21:01:15,795][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 21:01:17,683][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 21:01:19,476][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 21:01:21,268][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 21:01:21,270][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 21:01:21,502][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_2 - 5_
[ 2019-04-23 21:01:21,729][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_2 - 5_
[ 2019-04-23 21:01:22,070][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_2 - 5_
[ 2019-04-23 21:01:22,292][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_2 - 5_
[ 2019-04-23 21:01:22,514][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_2 - 5_
[ 2019-04-23 21:01:22,515][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_2 - 5_
[ 2019-04-23 21:01:22,741][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_3 - 5_
[ 2019-04-23 21:01:22,960][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_3 - 5_
[ 2019-04-23 21:01:23,183][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_3 - 5_
[ 2019-04-23 21:01:23,509][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_3 - 5_
[ 2019-04-23 21:01:23,731][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_3 - 5_
[ 2019-04-23 21:01:23,732][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_3 - 5_
[ 2019-04-23 21:01:23,733][cascade_classifier.calc_accuracy] Accuracy(layer_2 - train.classifi
[ 2019-04-23 21:01:23,919][cascade_classifier.fit_transform] [layer=3] look_indexs=[0], X_cur_
```

```
[ 2019-04-23 21:01:25,814][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 21:01:27,814][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 21:01:29,804][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 21:01:31,581][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 21:01:33,575][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 21:01:33,579][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 21:01:35,461][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 21:01:37,341][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 21:01:39,229][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 21:01:41,114][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 21:01:43,103][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 21:01:43,104][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 21:01:43,330][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_2 - 5_
[ 2019-04-23 21:01:43,552][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_2 - 5_
[ 2019-04-23 21:01:43,776][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_2 - 5_
[ 2019-04-23 21:01:43,998][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_2 - 5_
[ 2019-04-23 21:01:44,221][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_2 - 5_
[ 2019-04-23 21:01:44,222][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_2 - 5_
[ 2019-04-23 21:01:44,442][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_3 - 5_
[ 2019-04-23 21:01:44,663][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_3 - 5_
[ 2019-04-23 21:01:44,883][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_3 - 5_
[ 2019-04-23 21:01:45,103][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_3 - 5_
[ 2019-04-23 21:01:45,322][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_3 - 5_
[ 2019-04-23 21:01:45,323][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_3 - 5_
[ 2019-04-23 21:01:45,325][cascade_classifier.calc_accuracy] Accuracy(layer_3 - train.classifi
[ 2019-04-23 21:01:45,509][cascade_classifier.fit_transform] [layer=4] look_indexs=[0], X_cur_
[ 2019-04-23 21:01:47,608][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 21:01:49,600][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 21:01:51,584][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 21:01:53,370][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 21:01:55,457][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 21:01:55,458][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 21:01:57,349][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 21:01:59,229][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 21:02:01,229][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 21:02:03,319][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 21:02:05,304][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 21:02:05,306][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 21:02:05,525][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_2 - 5_
[ 2019-04-23 21:02:05,744][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_2 - 5_
[ 2019-04-23 21:02:05,966][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_2 - 5_
[ 2019-04-23 21:02:06,189][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_2 - 5_
[ 2019-04-23 21:02:06,409][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_2 - 5_
[ 2019-04-23 21:02:06,410][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_2 - 5_
[ 2019-04-23 21:02:06,635][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_3 - 5_
[ 2019-04-23 21:02:06,857][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_3 - 5_
[ 2019-04-23 21:02:07,073][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_3 - 5_
[ 2019-04-23 21:02:07,294][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_3 - 5_
```

```
[ 2019-04-23 21:02:07,511][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_3 - 5_
[ 2019-04-23 21:02:07,512][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_3 - 5_
[ 2019-04-23 21:02:07,514][cascade_classifier.calc_accuracy] Accuracy(layer_4 - train.classifi
[ 2019-04-23 21:02:07,515][cascade_classifier.fit_transform] [Result][Optimal Level Detected]
```

In [46]:
```
print('Training complete in {:.0f}m {:.4f}s'.format(
        time_elapsed // 60, time_elapsed % 60))
```

```
Training complete in 2m 5.3713s
```

In [47]:
```
pred = predict_labels(gc,features_test.toarray())
pred_scores.append(("DCF", [precision_score(labels_test,pred), recall_score(labels_tes
```

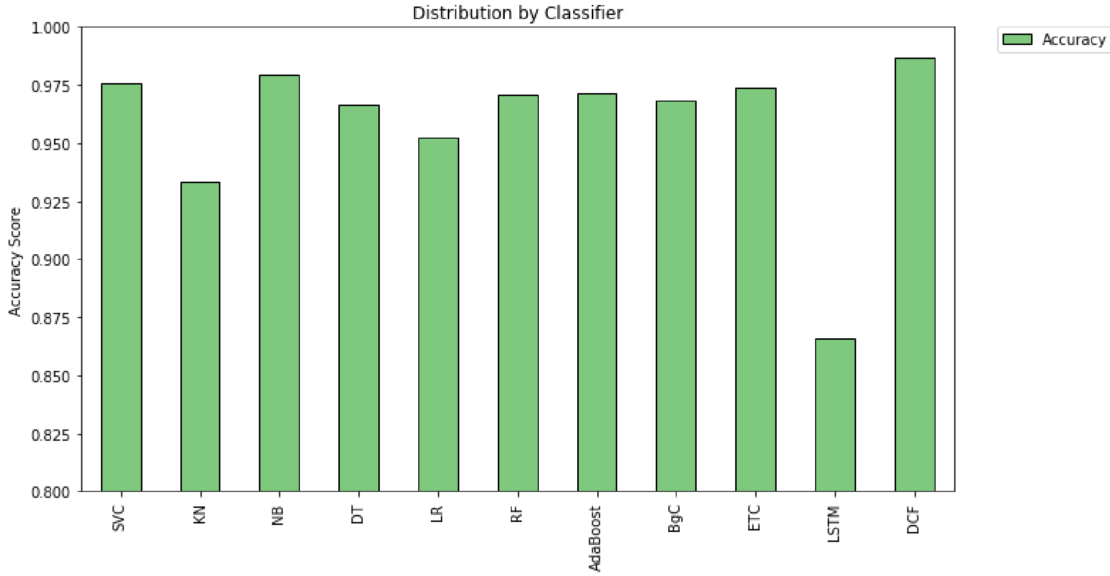
```
[ 2019-04-23 21:02:07,611][cascade_classifier.transform] X_groups_test.shape=[(1672, 9403)]
[ 2019-04-23 21:02:07,740][cascade_classifier.transform] group_dims=[9403]
[ 2019-04-23 21:02:07,740][cascade_classifier.transform] X_test.shape=(1672, 9403)
[ 2019-04-23 21:02:07,818][cascade_classifier.transform] [layer=0] look_indexs=[0], X_cur_test
[ 2019-04-23 21:02:09,504][cascade_classifier.transform] [layer=1] look_indexs=[0], X_cur_test
```

In [48]:
```
df = pd.DataFrame.from_items(pred_scores,orient='index', columns=['Precision', 'Recal
df
```

Out[48]:

	Precision	Recall	Accuracy	F1	Training Time (s)
SVC	0.989305	0.825893	0.975478	0.900243	0m 0.3595s
KN	0.991304	0.508929	0.933612	0.672566	0m 0.0031s
NB	0.939535	0.901786	0.979067	0.920273	0m 0.0019s
DT	0.878378	0.870536	0.966507	0.874439	0m 0.2016s
LR	0.967532	0.665179	0.952153	0.788360	0m 0.0097s
RF	1.000000	0.781250	0.970694	0.877193	0m 1.2949s
AdaBoost	0.948980	0.830357	0.971292	0.885714	0m 2.5555s
BgC	0.905213	0.852679	0.968301	0.878161	0m 1.0332s
ETC	0.978723	0.821429	0.973684	0.893204	0m 0.9084s
LSTM	0.000000	0.000000	0.866029	0.000000	9m 49.5561s
DCF	0.980952	0.919643	0.986842	0.949309	2m 5.3713s

In [49]:
```
df.plot(kind='bar', y="Accuracy", ylim=(0.8,1.0), figsize=(11,6), align='center', colo
plt.xticks(np.arange(11), df.index)
plt.ylabel('Accuracy Score')
plt.title('Distribution by Classifier')
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.savefig("../img/sms-acc-v1-stop.eps")
plt.show()
```




```
In [50]: import pickle
         # dump
         with open("../pkl/sms-gc-v1-stop.pkl", "wb") as f:
             pickle.dump(gc, f, pickle.HIGHEST_PROTOCOL)

         # # load
         # with open("../pkl/2018_gc.pkl", "rb") as f:
         #     gc = pickle.load(f)
```

0.0.7 Final verdict - gcForest is your friend in spam detection.

```
In [ ]:
```