Prediction of Duplicate Bug Reports

Presented by : Mayuresh Nene, Prasad Chavan, Ryan Chui

Agenda

- Problem definition
- Datasets
- Solution
- Preprocessing
- Results
- Validation
- Limitations

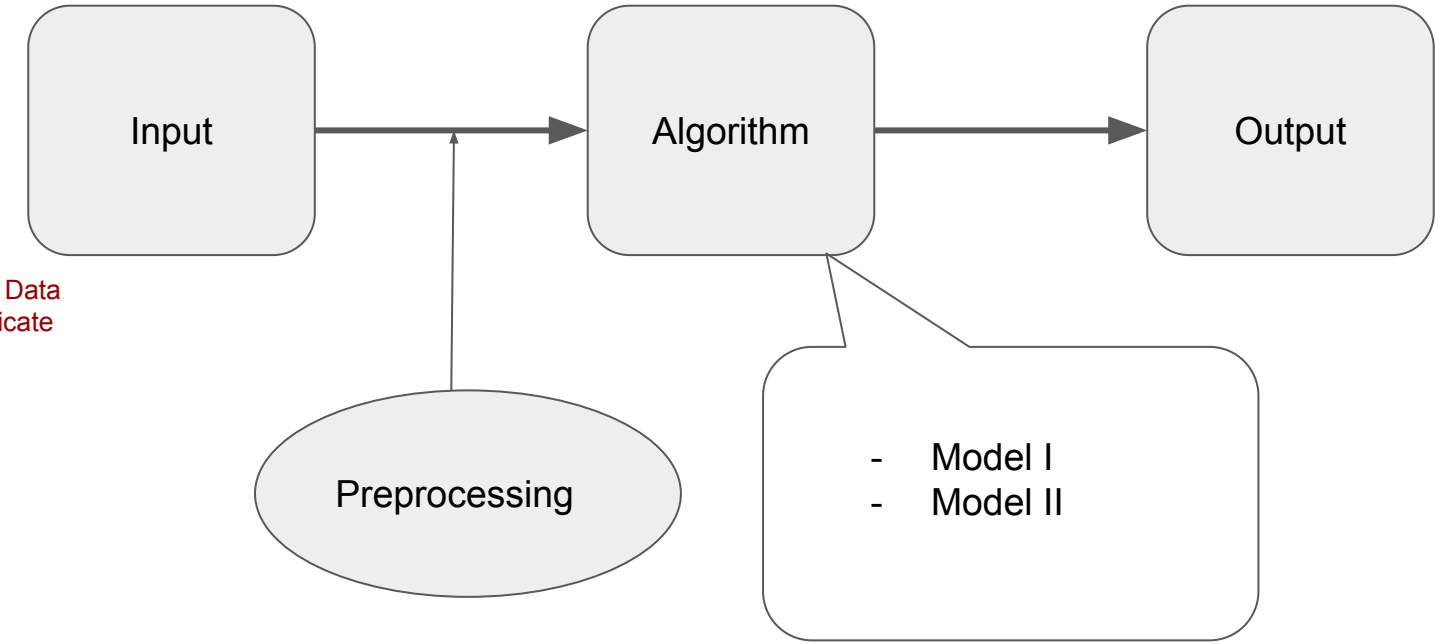
Problem Definition

- Bug reporting and fixing is a continuous process.
- Fixing bugs is a time-consuming process.
- Multiple users might report the same bug unknowingly.
- Developers spend 75% of their time fixing bugs.
- Increases inefficiencies.

Datasets

- Three different datasets -
  - Eclipse contains 46910 data points (
    - 34,223 Non-duplicates and 12,668 Duplicates).
  - Mozilla contains 60906 data points
    - (36,834 Non-duplicates and 24,072 Duplicates).
  - Thunderbird contains 14265 data points
    - (9906 Non-duplicates and 4359 Duplicates).
- Combined total of 122,441 data points (80,963 Non-duplicates and 41,099).
- We merged all three datasets into one huge dataset and ran experiments on it.

Solution to the Problem



Before Preprocessing

in the example page see url i have tried to vertically align all the cells using col valigntop mac mozilla . and win firefox . both demonstrate this problem so it seems to be quite an old issue. it might be related to bug though i dont believe that dealt with valign. the page httptelcontar.netmiscplugins also shows that col wont apply background colours to cells either as well as not supporting valign the whole left column is meant to be green but mozfirefox wont display the colour this is possibly bug . i have tried to use css instead using verticalalign instead of valign i tried applying styleverticalalign top to the col tags as a workaround in case this issue is due to a lack of support for outdated html . properties and this did not work either col seems to not apply any formatting to cells whatsoever neither css nor html .. microsoft ie .. for the macintosh also displays the valign issue no support for either form of vertical alignment in col but icab b mac browser handles valign correctly not tried it with verticalalign as its css support is lacking. i assume that there is some workaround using css styles applied to table cells but i cannot get any styles to apply to the middle table cells alas its beyond me at the moment. but in case i do thats why ive preserved indexold.php . open page in mozilla or firefox the page doesnt look right. lack of top vertical alignment and lack of other formatting in the centre content table made it look right. all columns should be top verticallyaligned

Preprocessing: Tokenization

```
def identify_tokens(row):
    review = row['Description1']
    tokens = nltk.word_tokenize(review)
    token_words = [w for w in tokens if w.isalpha()]
    return token_words

data_df['Description1_clean'] = data_df.apply(identify_tokens, axis=1)
nodup_df['Description1_clean'] = nodup_df.apply(identify_tokens, axis=1)

def identify_tokens(row):
    review = row['Description2']
    tokens = nltk.word_tokenize(review)
    token_words = [w for w in tokens if w.isalpha()]
    return token_words
```

['in', 'the', 'example', 'page', 'see', 'url', 'i', 'have', 'tried', 'to', 'vertically', 'align', 'all', 'the', 'cells', 'using', 'col', 'valigntop', 'mac', 'seems', 'to', 'be', 'quite', 'an', 'old', 'is', 'dealt', 'with', 'valign', 'the', 'page', 'also', 'as', 'well', 'as', 'not','supporting', 'firefox', 'wont', 'display', 'the', 'col', 'ng', 'verticalalign', 'instead', 'of', 'valig', 'workaround', 'in', 'case', 'this','issue', 'nd', 'this','did', 'not', 'work', 'either', 'col', 'er', 'css', 'nor', 'html', 'microsoft', 'ie','for', 'either', 'form', 'of', 'vertical', 'alignment', 'in', 'col', 'but', 'icab', 'b', 'mac', 'browser', 'handles', 'valign', 'correctly', 'not', 'tried','it','with', 'verticalalign', 'as', 'its', 'css', 'support', 'is', 'lacking','i', 'assume', 'that', 'there', 'is','some', 'workaround', 'using', 'css', 'styles','applied', 'to', 'table', 'cells', 'but', 'i', 'can', 'not', 'get','any', 'styles', 'to', 'apply', 'to', 'the', 'middle', 'table', 'cells', 'alas', 'its', 'beyond', 'me', 'at', 'the','moment', 'but', 'in', 'case', 'i', 'do','thats', 'why', 'ive', 'preserved', 'open', 'page', 'in', 'mozilla','or', 'firefox','the', 'page', 'doesnt', 'look', 'right', 'lack', 'of', 'top', 'vertical', 'alignment', 'and','lack', 'of', 'other', 'formatting', 'in', 'the', 'centre', 'content', 'table', 'made', 'it', 'look', 'right', 'all','columns', 'should', 'be', 'top', 'verticallyaligned']

Preprocessing: Stop Word Removal

['example', 'page', 'see', 'url', 'tried', 'vertically', 'align', 'cells', 'using', 'col', 'valigntop','mac','mozilla', 'win', 'firefox',
'demonstrate', 'pro
'valign', 'page', 'als
'whole', 'left','colun
'instead', 'using', 'v
'case', 'issue', 'due
'cells', 'whatsoeve
'either', 'form', 'ver
'verticalalign', 'css'
'styles', 'apply', 'mi
'mozilla','firefox', 'p
'table', 'made', 'loo

```
def remove_stops(row):
    my_list = row['Description1_clean']
    meaningful_words = [w for w in my_list if not w in stops]
    return (meaningful_words)

data_df['Description1_clean'] = data_df.apply(remove_stops, axis=1)
nodup_df['Description1_clean'] = nodup_df.apply(remove_stops, axis=1)

def remove_stops(row):
    my_list = row['Description2_clean']
    meaningful_words = [w for w in my_list if not w in stops]
    return (meaningful_words)

data_df['Description2_clean'] = data_df.apply(remove_stops, axis=1)
nodup_df['Description2_clean'] = nodup_df.apply(remove_stops, axis=1)
```

nt', 'believe', 'dealt',
vell', 'supporting', 'valign',
ug', 'tried', 'use','css',
, 'col', 'tags', 'workaround',
ems', 'apply', 'formatting',
lign', 'issue','support',
rectly', 'tried',
ed', 'table', 'cells', 'get',
erved', 'open', 'page',
matting','centre', 'content',

Preprocessing: Lemmatization

example page see url tried vertical align cells using col valigntop mac mozilla win firefox demonstrate problem seem quite old issue might related bug though dont believe dealt valign page also shows col wont apply background colours cells either well support valign whole left column meant green mozfirefox wont display colour possibly bug tried use css instead verticalalign instead valign tried apply styleverticalalign top col tags workaround case issue due lack support outdated html properties work either col seems apply formatting cells whatsoever neither css html microsoft ie macintosh also displays valign issue support either form vertical alignment col icab b mac browser handles valign correctly tried verticalalign css support lacking assume workaround using css styles applied table cells get styles apply middle table cells alas beyond moment case thats ive preserved open page mozilla firefox page doesnt look right lack top vertical alignment lack formatting centre content table made look right columns top verticallyaligned

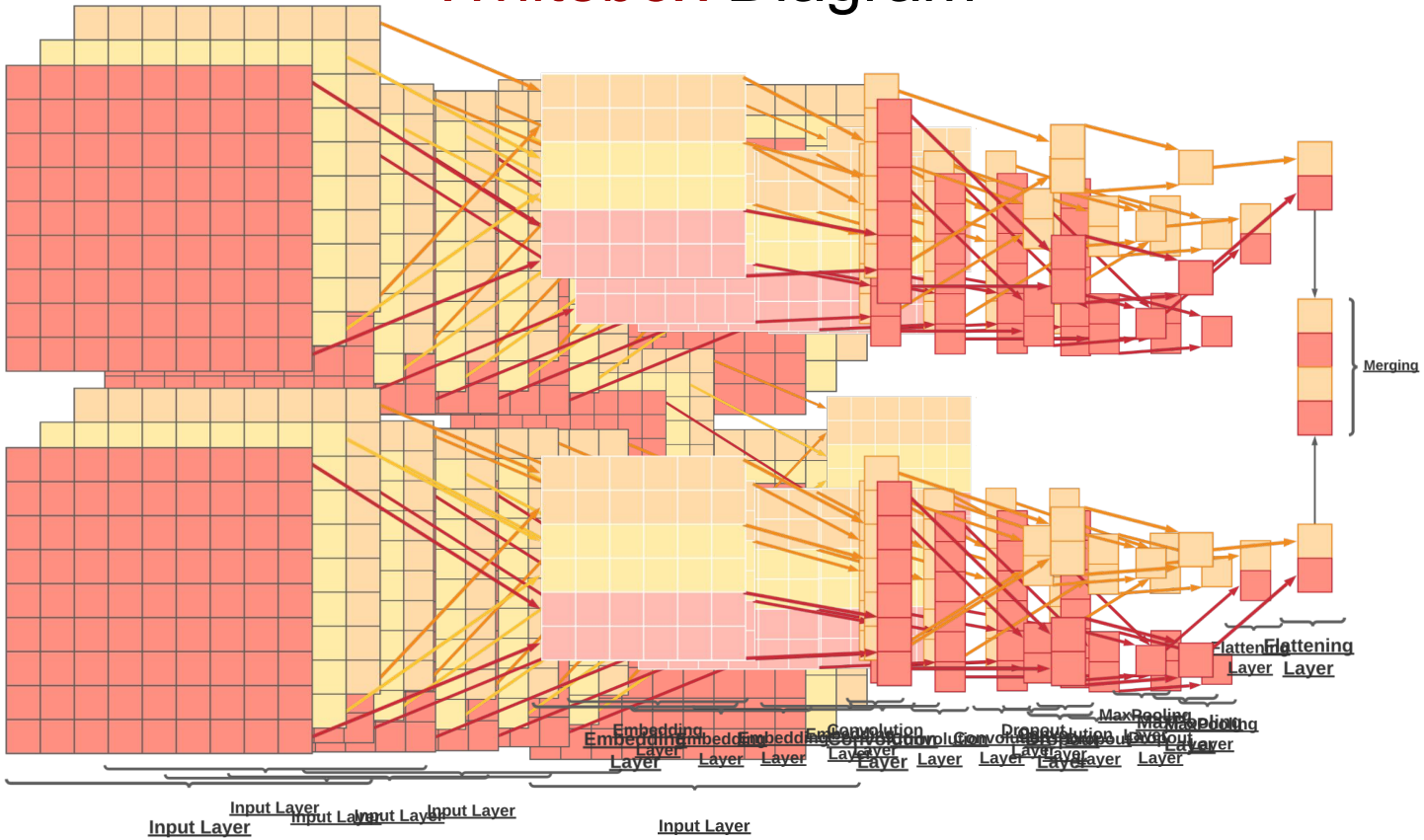
Preprocessing: Combining the Title and Description back

Once all the preprocessing is complete, the Title_1, Title_2, Description_1 and Description_2 are combined back together and their label is displayed beside them

This forms the basis of the input layer for the multichannel CNN that is to be deployed

	Processed	Label
0	dialup properties profile exposed prefs panels...	0
1	would really like plugin manager browser allow...	0
2	language encodings listed seemingly random ord...	0
3	using synaptics touch pad latest win driver v ...	0
4	history window select scroll current site open...	0
5	state checkbox dont allow removed cookies acce...	0
6	build mac os reproduce open history window loo...	0
7	since days search bookmarks gone search menu b...	0
8	mozilla save respond posted form correctly ins...	0
9	rightclicking back button list appears previou...	0
10	localization problems bookmarks sorted menu hi...	0
11	update quicktime solve bug similar cases need ...	0
12	rfe mozilla support x session management see x...	0
13	trying look form history autocomplete accident...	0
14	bookmark pointing host responding either shut ...	0

Whitebox Diagram



Whitebox Diagram

Multichannel CNN is deployed using keras

We made the model run through 3 epochs

Each epoch took an average of 50 minutes when ran multiple times

```
Max document length: 1734
Vocabulary size: 192248
(235065, 1734) (58767, 1734)
(235065,) (235065,)
Model: "model_2"
```

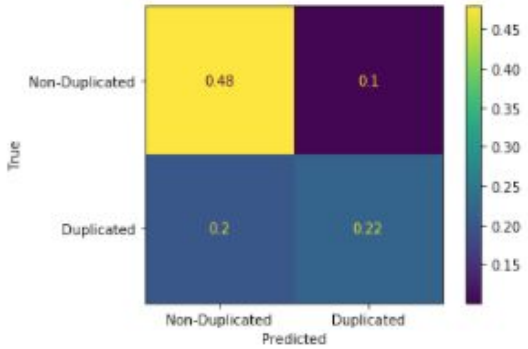
Layer (type)	Output Shape	Param #	Connected to
input_6 (InputLayer)	[(None, 1734)]	0	
input_7 (InputLayer)	[(None, 1734)]	0	
input_8 (InputLayer)	[(None, 1734)]	0	
embedding_5 (Embedding)	(None, 1734, 100)	19224800	input_6[0][0]
embedding_6 (Embedding)	(None, 1734, 100)	19224800	input_7[0][0]
embedding_7 (Embedding)	(None, 1734, 100)	19224800	input_8[0][0]
conv1d_5 (Conv1D)	(None, 1731, 32)	12832	embedding_5[0][0]
conv1d_6 (Conv1D)	(None, 1729, 32)	19232	embedding_6[0][0]
conv1d_7 (Conv1D)	(None, 1727, 32)	25632	embedding_7[0][0]
dropout_5 (Dropout)	(None, 1731, 32)	0	conv1d_5[0][0]
dropout_6 (Dropout)	(None, 1729, 32)	0	conv1d_6[0][0]
dropout_7 (Dropout)	(None, 1727, 32)	0	conv1d_7[0][0]
max_pooling1d_5 (MaxPooling1D)	(None, 865, 32)	0	dropout_5[0][0]
max_pooling1d_6 (MaxPooling1D)	(None, 864, 32)	0	dropout_6[0][0]
max_pooling1d_7 (MaxPooling1D)	(None, 863, 32)	0	dropout_7[0][0]
flatten_5 (Flatten)	(None, 27680)	0	max_pooling1d_5[0][0]
flatten_6 (Flatten)	(None, 27648)	0	max_pooling1d_6[0][0]
flatten_7 (Flatten)	(None, 27616)	0	max_pooling1d_7[0][0]
concatenate_2 (Concatenate)	(None, 82944)	0	flatten_5[0][0] flatten_6[0][0] flatten_7[0][0]
dense_4 (Dense)	(None, 10)	829450	concatenate_2[0][0]
dense_5 (Dense)	(None, 1)	11	dense_4[0][0]

Validation

How is the Logistic Regression performed for Title & Description fields for all three datasets?

	precision	recall	f1-score	support
0	0.70	0.83	0.76	45027
1	0.68	0.52	0.59	32546
accuracy			0.70	77573
macro avg	0.69	0.67	0.68	77573
weighted avg	0.70	0.70	0.69	77573

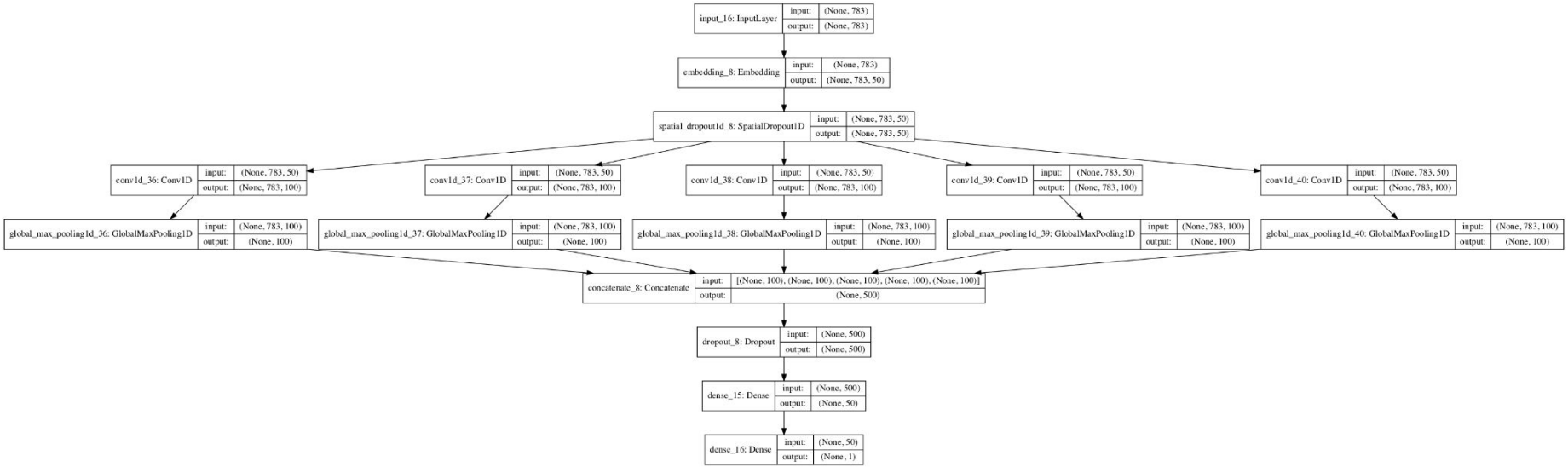
```
[[37220  7807]
 [15658 16888]]
```



TP = 0.22 suggested that the algorithm predict that Title & Description fields are correctly label as duplicated and they actually are duplicate.
TN = 0.48 suggested that the algorithm predicted non-duplicate and they are actually not duplicate.
FP (Type I Error) suggested that when Title & Description fields are actually non-duplicated, there is 0.1 probability that it predicts as duplicated.
FN (Type II Error) suggested that when Title & Description fields are duplicated, there is a 0.2 probability that it predicts as non-duplicate. (More Serious!)

Validation

How is our Multi Channel model defined?



- Input layer that defines the length of input sequences.
- Embedding layer set to the size of the vocabulary and 100-dimensional real-valued representations.
- One-dimensional convolutional layer with 32 filters and a kernel size set to the number of words to read at once.
- Max Pooling layer has consolidated and flatten layer reduce from three-dimensional to two dimensional output.

Validation

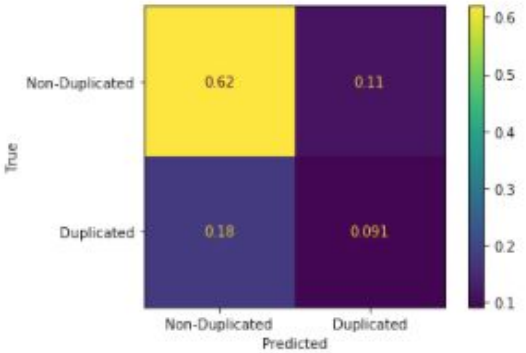
Performance of MC-CNN on ThunderBird and Eclipse Dataset (Include Title and Description Fields)

```
[[23250  4224]
 [ 6631  3420]]
              precision    recall  f1-score   support

         0.0       0.78      0.85      0.81     27474
         1.0       0.45      0.34      0.39     10051

    accuracy                           0.71     37525
   macro avg       0.61      0.59      0.60     37525
weighted avg       0.69      0.71      0.70     37525

Training Accuracy: 0.8546
Testing Accuracy:  0.7107
```



Eclipse

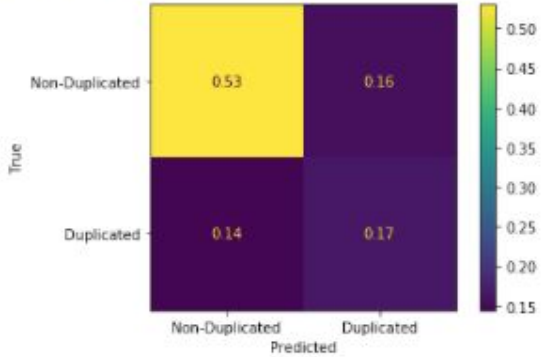
```
[[6050 1805]
 [1652 1904]]
              precision    recall  f1-score   support

           0       0.79      0.77      0.78      7855
           1       0.51      0.54      0.52      3556

    accuracy                           0.70     11411
   macro avg       0.65      0.65      0.65     11411
weighted avg       0.70      0.70      0.70     11411

Training Accuracy: 0.8860
Testing Accuracy:  0.6970

[Text(0, 0.5, 'True'), Text(0.5, 0, 'Predicted')]
```



ThunderBird

True Positive (Duplicate):

Issue Number: 75927

outlook express has ts a button that has the same effect as file compact ts folder. inclusion of ts feature would bring the mark as deleted mode to a fully featured level for users that work off grapcal elements as opposed to m or line commands.

After Stop word Removal: outlook express ts button effect file compact ts folder inclusion ts feature would bring mark deleted mode fully featured level users work grapcal elements opposed line commands

Issue Number: 242959

thunderbird supports saving and opening .eml files through the thunderbird interface but it should also support opening files externally. if thunderbird is my default mail client .eml files and maybe .msg too should be associated with and open in thunderbird. i would like to be able to doubleclick on a .eml file in explorer and have the file open in thunderbird instead of outlook express. . save an email as a .eml file. find the file in explorer. doubleclick on the file the file opens in outlook express if at all. the file should open in thunderbird just like if i had opened it from the menu.

After Stop word Removal: thunderbird supports saving opening files thunderbird interface also support opening files externally thunderbird default mail client files maybe associated open thunderbird would like able doubleclick file explorer file open thunderbird instead outlook express save email file find file explorer doubleclick file file opens outlook express file open thunderbird like opened menu

The actual value was positive and the model predicted label '1' value.

True Negative (Non-Duplicate):

Issue Number: 259628

the following html code is not shown correctly opera and ie show an other result table width border colgroup col alignleft col aligncenter col alignright colgroup tr tdlefttd tdcentertd tdrighttd tr table all tabledatas are aligned left align table data . left . center . right

After Stop word removal:

following html code shown correctly opera ie show result table width border colgroup col alignleft col aligncenter col alignright colgroup tr tdlefttd tdcentertd tdrighttd tr table tabledatas aligned left align table data left center right

Issue Number: 412748

the html col tag does not process all valid attributes. the similar bugs reported are from and earlier. the col is suppose to modify a table display. only some attributes work. there is an example at httpwww.wschools.comtagstryit.aspfilenametryhtmlcoltest which illustrates the problem note nothing is right justified. however some attributes such as width are processed. text should be right justified in column three but is not. it works in opera and ie it fails in chrome safari for windows

After Stop word Removal:

html col tag process valid attributes similar bugs reported earlier col suppose modify table display attributes work example illustrates problem note nothing right justified however attributes width processed text right justified column three works opera ie fails chrome safari windows

The actual value was 0 and the model predicted label '0' value.

False Positive:'

Issue Number: 421292
we should refer to the icons from org.eclipse.ui.images in our platform ui bundles.
see bug
After Stop word removal :
org.eclipse.ui, images, bug platform

Issue number: 422040
i open this bug for the work to use the new images provided by
org.eclipse.ui.images.
After Stop word removal :
Bug org.eclipse.ui, images

The actual value was 0 and the model predicted label '1' value.
(Type I Error)

False Negative:

Issue number: 344692
firefox ... is not closing down correctly. when i use the windows x to exit fx it exits the fx window but leaves fx running in the background. when i use the file menu and choose exit. i get the same result i am forced to ctrlaltdel to end the process. so i can restart fx. . open fx . use the windows x to close fx . fx window but leaves fx running in the background. . open fx . use the file menu and choose exit. . same as step . closes the fx window but leaves fx running in the background fx should exit leaving no process running in background. extensions talkback ... ie tab .. clicktab .. themes firefox default .

Issue Number: 246942
downloaded the browser rebooted and the program will not launch. no error messages upon installation. i had turned off my antivirus software while installing. after trying to open i can see the process started however the application never launches. .click on desktop icon .start program from task manager .open firefox.exe from program file i get an hour glass for a couple of seconds then absolutely nothing launched the program

The actual value was duplicated and the model predicted label '0' value.
(Type II Error)

Limitations and Future Improvements

- Can prove to be expensive as there is a lot of computation time.
- Feature selection can prove to be a bottleneck.
- May need additional features in case of score saturation.
- Solve the class imbalance problem in the data.