Preprocessing steps :

1. White space removal
2. Tokenization
3. Stop words removal
4. Punctuation removal
5. Url Removal
6. Spelling Correction
7. Lemmatization
8. Adding <s> at the start of tweet and <\s> at the end

Methodology :

Smoothing algorithm used: Laplace smoothing

ML Model Used for extrinsic evaluation: Random forest classifier with 13 estimators

Calculation of Beta : We created two language models, one for positive bigrams and another for negative bigrams. In order to generate sentimentally “louder” sentences(sentences with large sentiment component), we selectively increased the counts of those bigrams which appeared in the positive tweets for the positive LM and negative tweets for the negative LM proportionately with their frequency of occurance in the positive and negative corpus respectively. We then normalised it so that our model still stores probabilities.

We tweaked the bigram probability formula with laplace smoothing to come up with the following formula for calculating

bigram probabilities and used it to generate sentimentally “louder” sentences.

$$P(w(i) \mid w(i-1)) = \frac{Count(w(i-1)w(i)) + 1 + k \times Count^{*}(w(i-1)w(i))}{Count(w(i-1)) + V + Count^{*}(w(i-1))}$$

$P(w(i) \mid w(i-1)) \rightarrow$ Bigram probability of w(i-1)w(i)

$Count(w(i-1)w(i)) \rightarrow$ Count of the bigram $w(i-1)w(i)$ in the original corpus

$Count^{*}(w(i-1)w(i)) \rightarrow$ Count of the bigram $w(i-1)w(i)$ in the corpus with exclusively positive/negative sentences respectively.

$Count(w(i-1)) \rightarrow$ Count of unigram w(i-1) in the original corpus

$Count^{*}(w(i-1) \rightarrow$ Count of unigram w(i-1) in the corpus with exclusively positive/negative sentences respectively.

V $\rightarrow$ unique words in the original corpus

For this, we segregated the original corpus into two corpora, one with exclusively positive labels and the other with negative labels. Then we created dictionaries of unigrams vs unigram counts and bigrams vs bigram counts for the respective corpora, required for the calculation of bigram probabilities using the above-modified formula.

The above solution also punishes the bigrams not occurring in the positive/negative corpus by increasing the denominator $(Count * (w(i-1)))$ while not increasing the numerator.

Helper Functions:

1. Preprocess : Used to preprocess the given dataset as well as the test dataset. Accepts the dataframe with unprocessed text and outputs a dataframe with processed text.
2. Pickel: Used to save/retirieve bigram/Unigram models. Pickel.dump Inputs a python object and saves it on the disk. Alternatively, Pickel.load loads the saved python object saved in the disk
3. A function to calculate perplexity of the given sentence.

Outputs:

1. Average Perplexity of 500 sentences

```
average perplexity score of 500 generated sentences :  3391.67
```

2. Top-4 bigrams and their score after smoothing.

```
top 4 bigrams
(('http', '</s>'), 0.014073287307488051)
(('lol', '</s>'), 0.008906021533962515)
(('gon', 'na'), 0.0082332298558105)
(('day', '</s>'), 0.006479767257339328)
```

3. Accuracy of test set using dataset A and B for training.

```
accracy on Dataset A is : 0.785714285714285
accracy on dataset B is : 0.796583850931677
```

4. 10 generated samples:
   5 positives

```
<s> good severe marleematlin kreesha mind night nj deeper toast list geog yasmimmm cameronreilly excitement backlog lord 3rd 1week suck harper </s>
<s> chopsuey2e zhang political fabuleuxdestin afterwards bad jeremy poke sowwiiiee tilde chick r2e2 yan dif super n areejo semuuaa afternoon pc uugggh </s>
<s> mosquito 4am lobster top vernongarrett calgary greatfitness knowwwwww mc insure shogi 5th benny doingwork l8ly popularity normally ncaa tweak costa </s>
<s> okay dunkndisorderly tutor mommapuff souvenir sketchbook mwahahaha financial goodnight jenny hoopinispassion ho wliet sudden kea34 sri join60seconds offline shaft secular matthewsheppard </s>
<s> lucypope hoaaaaaaaaaaaaa caffeine mark liturgy wheel ceiling toronto dow werewolfseth m0t0breath h0area getknif ed yourproxycomm product 45 affairs solar funky tent </s>
```

   5 negatives

```
<s> sometimes psp 14 kudos therealpickler handyman 1st whip strobelight sprinters wickets mya flint launch emilyybr owningg lisaworld kingdom domingo roxannegregorio doreenatdms </s>

<s> know transformers conscious italian natmcb78 chesterfield aurea officials cig muster brightside jajjaja friend elon drew123 cream mrstessyman mga springleaf anoopdoggdesai </s>

<s> festivallights ijustmightendit grizzly flight eu myinnersexygirl rotten demivenom ahmed nikkithebee cleftmommy0 217 indrairwan jk boyhood êμ msrnbjazz picture lad oxo 6 </s>

<s> diversitybgt ruthramirez diversitybgt hometown littlebites bonnaroo shrew fact academia knackered ergo clever u u tomsmithmcse nin morrison sl rt announcements alas </s>

<s> oh gallery sport grey district factory yeh american disneys pilot apply ì le cheese stem burn honorary realdeal 32 shallow section </s>
```