DOCUMENTATION FOR THE EVALUATION OF THE DEGREE OF PROFANITY IN THE TWITTER DATASET

IMPORTING THE DATASET

dataset = pd.read_csv("input_data.csv")

4	A	В	С	D	Е	F	G	Н	I	J	K	L	М	N	0	Р	Q	R	S	T	U
1	cc	ount	hate_spe	offensive	neither	class	tweet														
2	0	3	0	0	3	3	2 !!! RT @r	nayasolove	ly: As a wo	oman you s	houldn't c	complain ab	out cleani	ng up your	house. &a	amp; as a n	nan you sho	uld alway	s take the	trash out	
3	1	3	0	3	()						or cuffin dat									
4	2	3	0	_)		!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit													
5	3	3	_	_	-	l						k like a tranr	•								
6	4	6		_)						about me n	•		_						
7	5	3		_)					-	meclaim y					ut still fuck	ing with h	oes!	3514;€	514;&#
8	6	3			_)				-		TE on anoth		_		_					
9	7	3	_	_	_)						u big bitches	coming fo	or us skinn	y girls!!&#</td><td>8221;</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>10</td><td>8</td><td>3</td><td></td><td>_</td><td></td><td>)</td><td>1 " &</td><td>you might i</td><td>not get ya i</td><td>oitch back</td><td>& tha</td><td>its that "</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>-</td></tr><tr><td>11</td><td>9</td><td>3</td><td></td><td>_</td><td></td><td>)</td><td>_</td><td> 1:4-11-</td><td></td><td></td><td>-1111</td><td>d:</td><td></td><td>lilia Alaia Ca</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>12</td><td>10 11</td><td>3</td><td></td><td>_</td><td>_</td><td>)</td><td>1 " Murda</td><td></td><td></td><td>•</td><td>oi i waike</td><td>d into a conv</td><td>/ersation</td><td>like this. Sr</td><td>nn</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>13 14</td><td>12</td><td>3</td><td></td><td>_</td><td></td><td>l</td><td>1 "So hoes</td><td></td><td></td><td></td><td>go on IG</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>15</td><td>13</td><td>3</td><td></td><td></td><td></td><td>)</td><td>1 " bad bite</td><td></td><td></td><td></td><td>-</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>16</td><td>14</td><td>3</td><td></td><td>-</td><td>_</td><td>)</td><td>1 " bitch ge</td><td></td><td></td><td>that i like</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>17</td><td>15</td><td>3</td><td></td><td>_</td><td></td><td>)</td><td>1 " bitch ni</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>18</td><td>16</td><td>3</td><td></td><td>_</td><td></td><td>)</td><td>1 " bitch pl</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>19</td><td>17</td><td>3</td><td></td><td></td><td></td><td>)</td><td>1 " bitch w</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>20</td><td>18</td><td>3</td><td>0</td><td>3</td><td>(</td><td>)</td><td>1 " bitches</td><td></td><td></td><td>В"</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>21</td><td>19</td><td>3</td><td>0</td><td>3</td><td>(</td><td>)</td><td>1 " black be</td><td>_</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>22</td><td>20</td><td>3</td><td>0</td><td>3</td><td>(</td><td>)</td><td>1 " broke b</td><td>itch cant te</td><td>II me noth</td><td>ing"</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>G</td><td></td><td>Н</td><td></td><td></td><td>J</td><td></td><td>(</td><td>L</td><td>М</td><td>N</td><td></td><td>0</td><td>Р</td><td>Q</td><td></td><td>R</td><td>S</td><td>1</td><td>-</td><td>U</td><td>١</td></tr><tr><td>twee</td><td></td><td></td><td></td><td></td><td></td><td></td><td>•</td><td>_</td><td></td><td></td><td></td><td>_</td><td></td><td></td><td><u>'</u></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td>-</td><td></td><td></td><td></td><td></td><td></td><td>- lu</td><td>l-1l</td><td>4 -1</td><td></td><td>L</td><td> 0</td><td></td><td></td><td>-11-</td><td>l - l</td><td>4-1 41</td><td>4 l</td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td>•</td><td></td><td>•</td><td></td><td>n't comp</td><td></td><td></td><td></td><td></td><td>use. &am</td><td>p; as a i</td><td>man you</td><td>snould</td><td>always</td><td>take the</td><td>trasn c</td><td>out</td><td></td><td></td></tr><tr><td>!!!!!! F</td><td>₹T @m</td><td>leew1</td><td>.7: boy o</td><td>dats colo</td><td>ltyga</td><td>dwn b</td><td>ad for cuf</td><td>fin dat h</td><td>ioe in th</td><td>ie 1st pl</td><td>ace!!</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>!!!!!!</td><td>RT @</td><td>UrKind</td><td>OfBran</td><td>d Dawg</td><td>!!!! RT @</td><td>080sba</td><td>by4life:</td><td>You ever</td><td>fuck a l</td><td>bitch an</td><td>d she st</td><td>art to cry</td><td>? You b</td><td>e confu</td><td>sed as s</td><td>hit</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>look like</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td>•</td><td>_</td><td>_</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1 11 1</td><td>1</td><td>1</td><td>0.0570.6</td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td>_</td><td></td><td></td><td></td><td>•</td><td>near abou</td><td></td><td>-</td><td></td><td>_</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>!!!!!!</td><td></td><td>!!!!!"@</td><td>T_Madi</td><td>son_x:1</td><td>he shit</td><td>just b</td><td>ows me.</td><td>.claim yo</td><td>ou so fai</td><td>ithful an</td><td>id down</td><td>for som</td><td>ebody l</td><td>out still</td><td>fucking</td><td>with ho</td><td>es! </td><td>28514;&</td><td>#128514</td><td>1;ԅ</td><td>514;"</td></tr><tr><td>!!!!!!</td><td>'@B</td><td>righte</td><td>rDays: I</td><td>can not</td><td>just sit</td><td>up and</td><td>HATE or</td><td>n anothe</td><td>r bitch .</td><td>. I got to</td><td>o much</td><td>n shit goir</td><td>ng on!"</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>!!!!&i</td><td>#8220;</td><td>@selfi</td><td>ieaueer</td><td>nbri: cau</td><td>se I'm t</td><td>ired of</td><td>you big l</td><td>bitches o</td><td>oming</td><td>for us sk</td><td>inny gir</td><td>rls!!R</td><td>21:</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td>_</td><td>•</td><td></td><td></td><td></td><td>thats th</td><td></td><td></td><td> </td><td>7 0</td><td></td><td>7</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>o.com</td><td>יטע ,קוי</td><td>u iiiigi</td><td>it not ge</td><td>L ya Dit</td><td>UII DOCK</td><td>Gamp</td><td>, criacs tri</td><td>ut.</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></tbody></table>						

IMPORTING THE LIST OF BAD WORDS

```
text_file = open("badwords.txt", "rt")
```

If you look at the above dataset, then you will find the slur words and the usernames that we do not want for our analysis for profanity detection. There are many redundant words that we need to eliminate for easy computation.

Now, we will only work with column *tweet* and drop all the other columns as we don't need them for the analysis. Note that the column *tweet* will be at position B after removing the unwanted columns.

4	Α	В	С	D	E	F	G	Н	1	J	K	L	М	N	0	P	Q
1		tweet															
2	0	!!! RT @m	!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & amp; as a man you should always take the trash out														
3	1	!!!!! RT @	!!!!! RT @mleew17: boy dats coldtyga dwn bad for cuffin dat hoe in the 1st place!!														
4	2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit															
5	3	!!!!!!!!! RT	@C_G_An	derson: @	viva_base	d she look	like a tran	ny									
6	4	. !!!!!!!!!!!!	!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!														
7	5	!!!!!!!!!@T_Madison_x: The shit just blows meclaim you so faithful and down for somebody but still fucking with hoes! 😂 😂											514;€	514;"			
8	6	!!!!!!"@_	BrighterDa	ys: I can n	ot just sit (up and HAT	ΓE on anoth	ner bitch	I got too m	nuch shit go	ing on!"						
9	7	!!!!“);@selfieq	ueenbri: c	ause I'm ti	red of you	big bitche	s coming f	or us skinn	y girls!!	3221;						
10	8	" & y	ou might n	ot get ya b	itch back 8	& that	s that "										
11	9	11															
12	10	" Keeks is	a bitch she	curves ev	eryone " l	ol I walked	into a con	versation	like this. Sr	nh							
13	11	" Murda G	ang bitch it	ts Gang Lai	nd "												
14	12	" So hoes that smoke are losers ? " yea go on IG															
15	13	" bad bitches is the only thing that i like "															

CLEANING THE DATASET

We will now clean the data using clean text function.

After using this function, we will get the following output which is much easier to work with.

4	Α	В	C	D	Е	F	G	Н		J	K	L	M
1		tweet											
2	0	as a wor	nan you sh	ould not c	omplain ab	out cleani	ng up your	house am	o as a man	you shoul	d always ta	ke the tras	h out
3	1	boy dats	boy dats coldtyga dwn bad for cuffin dat hoe in the place										
4	2	dawg y	ou ever fu	ck a bitch a	nd she sta	to cry you	be confuse	d as shit					
5	3	she lool	k like a trar	nny									
6	4	the shit	the shit you hear about me might be true or it might be faker than the bitch who told it to ya										
7	5	the shit j	ust blows n	neclaim yo	u so faithf	ul and dov	n for some	body but	still fuckin	g with hoe	S		
8	6	i can not	just sit up a	and hate o	n another l	oitch i got	too much s	hit going o	n				
9	7	cause i an	n tired of y	ou big bitc	hes comin	g for us ski	nny						
10	8	amp you	might not (get ya bitcl	n back amp	thats that							
11	9	hobbies	include fig	hting mari	am bitch								
12	10	keeks is a	a bitch she	curves eve	ryone lol	i walked in	to a conve	rsation like	this smh				

STEMMING

In this, we will stem each word of a tweet.

```
def stem_words(getDataset):
    sentences = []
    for index, row in getDataset.iterrows():
        sentence = ''.join([snowball_stemmer.stem(word) for word in row["tweet"].split()])
        sentences.append(sentence)

return pd.DataFrame({"tweet": sentences})
```

4	Α	В	С	D	Е	F	G	Н	1	J	K	L	
1		tweet											
2	0	as a woma	as a woman you should not complain about clean up your hous amp as a man you should alway take the trash o										
3	1	boy dat co	boy dat coldtyga dwn bad for cuffin dat hoe in the place										
4	2	dawg you ever fuck a bitch and she sta to cri you be confus as shit											
5	3	she look l	ike a trann	i									
6	4	the shit yo	ou hear ab	out me mig	ght be true	or it migh	t be faker t	than the bi	tch who to	ld it to ya			
7	5	the shit ju	ist blow me	eclaim you	so faith ar	nd down fo	r someboo	di but still f	fuck with h	oe			
8	6	i can not j	ust sit up a	nd hate or	anoth bit	ch i got too	much shit	go on					
9	7	caus i am	tire of you	big bitch c	ome for us	skinni							
10	8	amp you r	night not g	et ya bitch	back amp	that that							

STOPWORDS

```
Then, we will remove the stopwords.
```

```
def remove_stop_words(getDataset):
    sentences = []
    for index, row in getDataset.iterrows():
        sentence = ' '.join([word for word in row["tweet"].split() if word not in all_stopwords])
        sentences.append(sentence)

return pd.DataFrame({"tweet": sentences})
```

Δ	Α	В	С	D	E	F							
1		tweet											
2	0	woman co	mplain cle	an hous ar	np man alv	way trash							
3	1	boy dat co	ooy dat coldtyga dwn bad cuffin dat hoe place										
4	2	dawg fuck	dawg fuck bitch sta cri confus shit										
5	3	look like t											
6	4	shit hear t	shit hear true faker bitch told ya										
7	5	shit blow	meclaim fa	ith someb	odi fuck ho	oe							
8	6	sit hate an	oth bitch g	got shit									
9	7	caus tire b	ig bitch co	me skinni									
10	8	amp ya bit	tch amp										
11	9	hobbi incl	ud fight m	ariam bitch	1								
12	10	keek bitch	curv ever	yon lol wal	k convers	like smh							

CENSORING THE SLUR WORDS AND CALCULATING THE DEGREE OF PROFANITY

In this, we will censor the slur words, using the list of bad words, and count the **number of slur** words per tweet as the degree of profanity.

THE FINAL RESULT

Then we will output the result in the *output.csv* file. # output the result temp.to_csv("output_data.csv")

1	Α	В	С	D
1		tweet	profanity_deg	gree
2	0	woman complain clean hous amp man alway trash	0	
3	1	boy dat coldtyga dwn bad cuffin dat hoe place	0	
4	2	dawg ***** **** sta cri confus *****	3	
5	3	look like tranni	0	
6	4	***** hear true faker **** told ya	2	
7	5	***** blow meclaim faith somebodi ***** hoe	2	
8	6	sit hate anoth ***** got *****	2	
9	7	caus tire big ***** come skinni	1	
10	8	amp ya ***** amp	1	