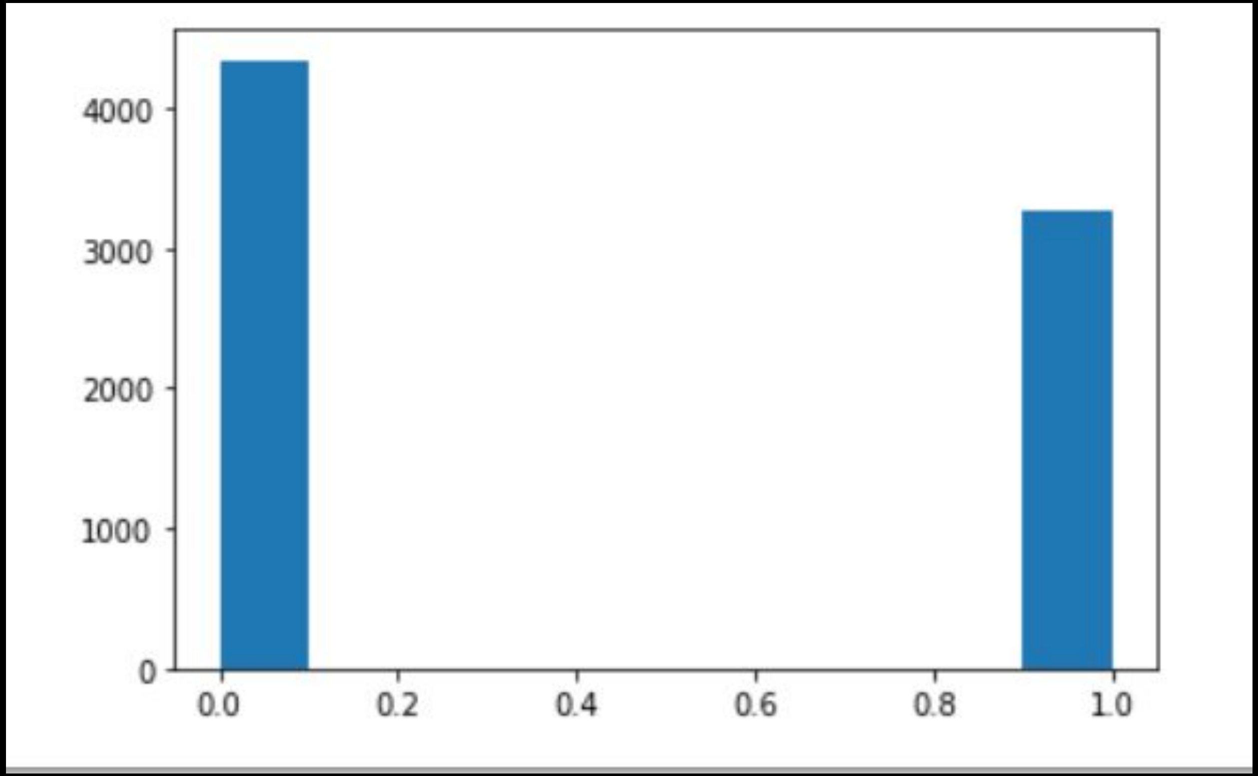
NAME = GAUTAM KUMAR
ROLL = B19EE031

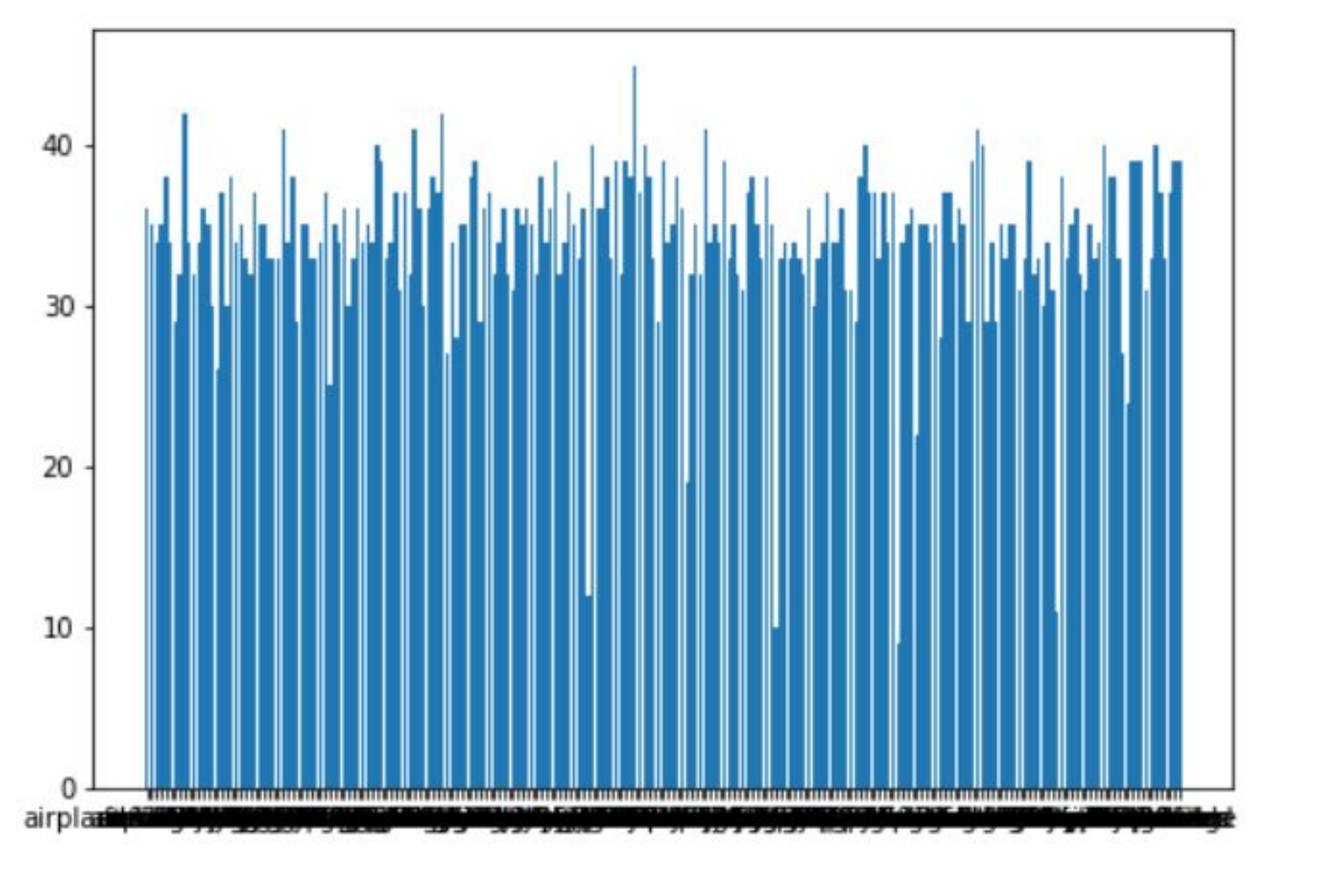
LAB 5 REPORT

TARGET COLUMN

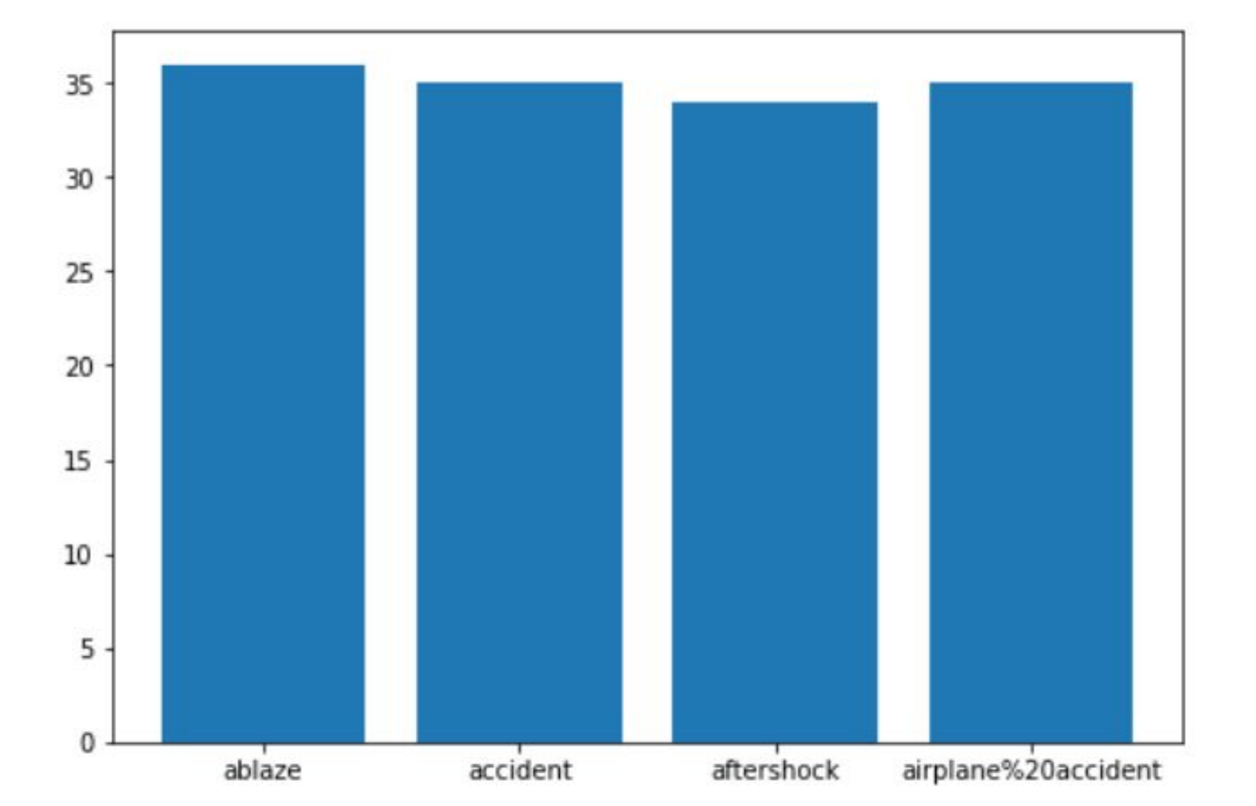
```
Counter({0:4342, 1:3271})
```



PLOT THE COUNT OF EACH KEYWORD



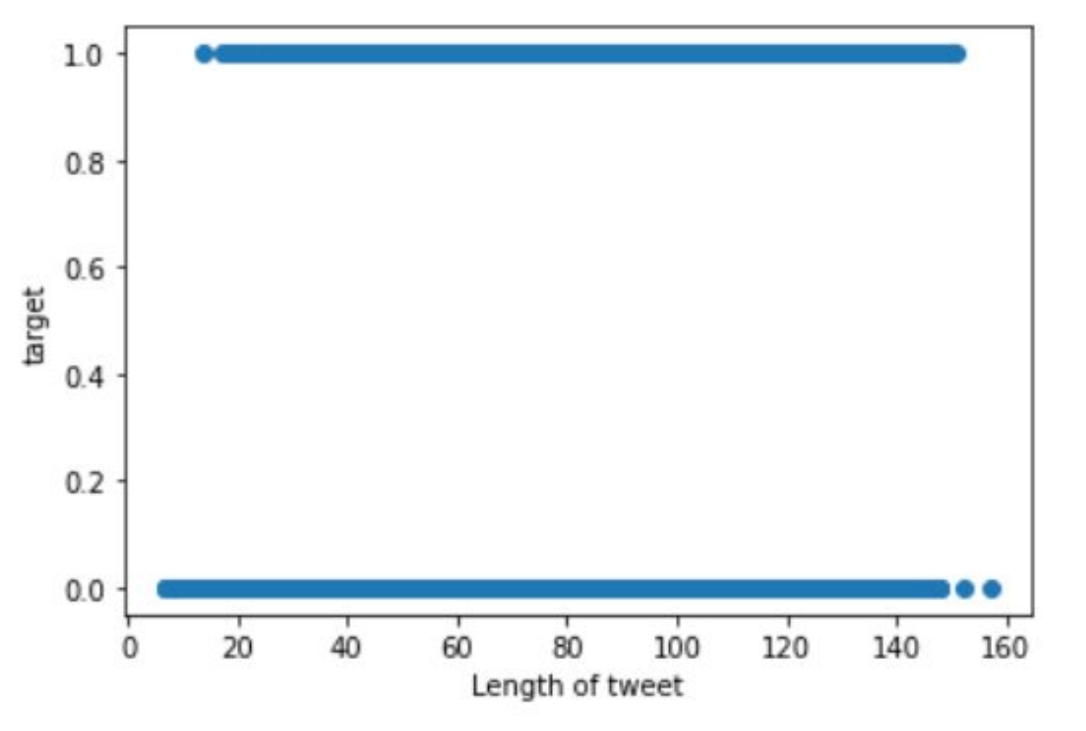
EXTRA ANALYSIS: PLOT OF COUNT OF FIRST 4 WORDS



RELATION BETWEEN LENGTH OF TWEET AND TARGET

They are positively correlated but the correlation ratio is small.

	len_tweet	target
len_tweet	1.000000	0.181817
target	0.181817	1.000000



NULL VALUES

```
id column has 0 NULL values
keyword column has 61 NULL values
location column has 2533 NULL values
text column has 0 NULL values
target column has 0 NULL values
len_tweet column has 0 NULL values
```

DATASET AFTER REMOVING EMOJI, LINKS, PUNCTUATION AND SPELLING CORRECTION

	id	keyword	location	text	target	len_tweet
31	48	ablaze	Birmingham	bbcmtd Wholesale Markets ablaze	1	55.0
32	49	ablaze	Est. September 2012 - Bristol	We always try to bring the heavy metal RT	0	67.0
33	50	ablaze	AFRICA	AFRICANBAZE Breaking newsNigeria flag set abla...	1	82.0
34	52	ablaze	Philadelphia, PA	Crying out for more Set me ablaze	0	34.0
35	53	ablaze	London, UK	On plus side LOOK AT THE SKY LAST NIGHT IT WAS...	0	76.0

EXTRA ANALYSIS: WORD CLOUD OF ENTIRE DATASET

WORD CLOUD OF REAL TARGET

WORD CLOUD OF FAKE TARGET

ONLY TEXT AND TARGET COLUMN PRESENT

	text	target
31	bbcmtd Wholesale Markets ablaze	1
32	We always try to bring the heavy metal RT	0
33	AFRICANBAZE Breaking newsNigeria flag set abla...	1
34	Crying out for more Set me ablaze	0
35	On plus side LOOK AT THE SKY LAST NIGHT IT WAS...	0

TDM OF ENTIRE DATASET

	0011	001116	005225	0104	010401	012032	012624	02	0215	03	0306	030811	034	0400	045	05	05082015	06	0605	061	06342
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5075	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5076	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5077	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5078	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5079	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

5080 rows × 13537 columns

TDM OF CLASS WITH TARGET = 1

	0011	001116	005225	0104	010401	012032	012624	030811	0400	05	05082015	06	061	063424	07	070	075	080	0800	0802pm
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2191	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2192	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2193	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2194	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2195	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2196 rows × 7241 columns

TDM OF CLASS WITH TARGET = 0

	02	0215	03	0306	034	045	05	06	0605	06jst	0700	0730	08072015	08315	0913	10	100	1000	100000	100mb	100nd	100s
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2879	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2880	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2881	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2882	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2883	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2884 rows × 9362 columns

SUM OF UNIQUE WORDS OF CLASS

```
Sum of unique words in class1 =  8861
Sum of unique words in class0 =  11518
Sum of unique words in both classes =  16844
```

So sum of unique words of class1 and class0 is not equal to sum of unique words in both classes.
Explanation: Since class1 contain words which is present in class0 also. For eg. Word ‘accident’ is present in both classes.
Hence the sum is more than no. of unique words in both classes

DATASET AFTER TRANSFORMATION

	0011	001116	005225	010401	012032	012624	02	0215	03	0306	030811	0400	045	05	06	0605	061	063424	06jst	07	070	07
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	(
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	(
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	(
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	(
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	(

5 rows × 11861 columns

CLASSIFICATION REPORT

	precision	recall	f1-score	support
class - 0	0.76	0.92	0.83	566
class - 1	0.86	0.64	0.74	450
accuracy			0.80	1016
macro avg	0.81	0.78	0.78	1016
weighted avg	0.81	0.80	0.79	1016

ANALYSIS

- Accuracy of test data = 0.7962598425196851

- Pipeline used : Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', MultinomialNB())])

- Confusion Matrix: array([[521, 45], [162, 288]])

CONCLUSION

- *Hence we learnt Natural Language Processing.*
- *We learnt to convert text to integers using Count Vectorizer.*
- *We learnt about TDM, TF, IDF transformations.*
- *Plotted the word cloud.*
- *Used pipeline of transformation.*
- *Used Multinomial model for analysis.*
- *Calculated precision, recall, f1 score and confusion matrix*