Fake News Detection

Importing Libraries and Data

In [64]:

```
# Import general useful packages
import numpy as np
import pandas as pd
import re

# Counter elements
from collections import Counter

# Matplot
import matplotlib.pyplot as plt
%matplotlib inline

# nltk
import nltk
from nltk.corpus import stopwords
from  nltk.stem import SnowballStemmer
from nltk.stem import WordNetLemmatizer #word stemmer class
lemma = WordNetLemmatizer()
from wordcloud import WordCloud, STOPWORDS
from nltk import FreqDist

# Import matplotlib for visualisations
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import seaborn as sns
import scikitplot as skplt

# Import all machine learning algorithms
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
import xgboost as xgb

# Import other useful subpackage
from sklearn.metrics import confusion_matrix, accuracy_score, classific
ation_report

import json
import os
import io

import warnings
warnings.filterwarnings("ignore")
%matplotlib inline
```

In [66]:

```
pip install wordcloud
```

```
Requirement already satisfied: wordcloud in c:\programdata\anaconda3\li
b\site-packages (1.5.0)
Requirement already satisfied: pillow in c:\programdata\anaconda3\lib\s
ite-packages (from wordcloud) (6.1.0)
Requirement already satisfied: numpy>=1.6.1 in c:\programdata\anaconda3
\lib\site-packages (from wordcloud) (1.16.4)
Note: you may need to restart the kernel to use updated packages.
```

In [67]:

```
pip install xgboost
```

```
Requirement already satisfied: xgboost in c:\programdata\anaconda3\lib
\site-packages (0.90)
Requirement already satisfied: numpy in c:\programdata\anaconda3\lib\si
te-packages (from xgboost) (1.16.4)
Requirement already satisfied: scipy in c:\programdata\anaconda3\lib\si
te-packages (from xgboost) (1.2.1)
Note: you may need to restart the kernel to use updated packages.
```

In [68]:

```
pip install scikit-plot
```

```
Requirement already satisfied: scikit-plot in c:\programdata\anaconda3
\lib\site-packages (0.3.7)
Requirement already satisfied: joblib>=0.10 in c:\programdata\anaconda3
\lib\site-packages (from scikit-plot) (0.13.2)
Requirement already satisfied: scipy>=0.9 in c:\programdata\anaconda3\l
ib\site-packages (from scikit-plot) (1.2.1)
Requirement already satisfied: matplotlib>=1.4.0 in c:\programdata\anac
onda3\lib\site-packages (from scikit-plot) (3.1.0)
Requirement already satisfied: scikit-learn>=0.18 in c:\programdata\ana
conda3\lib\site-packages (from scikit-plot) (0.21.2)
Requirement already satisfied: cycler>=0.10 in c:\programdata\anaconda3
\lib\site-packages (from matplotlib>=1.4.0->scikit-plot) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\programdata\anac
onda3\lib\site-packages (from matplotlib>=1.4.0->scikit-plot) (1.1.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1
in c:\programdata\anaconda3\lib\site-packages (from matplotlib>=1.4.0->
scikit-plot) (2.4.0)
Requirement already satisfied: python-dateutil>=2.1 in c:\programdata\a
naconda3\lib\site-packages (from matplotlib>=1.4.0->scikit-plot) (2.8.
0)
Requirement already satisfied: numpy>=1.11 in c:\programdata\anaconda3
\lib\site-packages (from matplotlib>=1.4.0->scikit-plot) (1.16.4)
Requirement already satisfied: six in c:\programdata\anaconda3\lib\site
-packages (from cycler>=0.10->matplotlib>=1.4.0->scikit-plot) (1.12.0)
Requirement already satisfied: setuptools in c:\programdata\anaconda3\l
ib\site-packages (from kiwisolver>=1.0.1->matplotlib>=1.4.0->scikit-plo
t) (41.0.1)
Note: you may need to restart the kernel to use updated packages.
```

In [38]:

```
# Reading the main train data file
df=pd.read_json("train.json")
df.head(20)
```

Out[38]:

	claim	claimant	date	id	label	related_articles

	claim	claimant	date	id	label	related_articles
0	A line from George Orwell's novel 1984 predict...		2017-07-17	0	0	[122094, 122580, 130685, 134765]
1	Maine legislature candidate Leslie Gibson insu...		2018-03-17	1	2	[106868, 127320, 128060]
2	A 17-year-old girl named Alyssa Carson is bein...		2018-07-18	4	1	[132130, 132132, 149722]
3	In 1988 author Roald Dahl penned an open lette...		2019-02-04	5	2	[123254, 123418, 127464]
4	When it comes to fighting terrorism, "Another ...	Hillary Clinton	2016-03-22	6	2	[41099, 89899, 72543, 82644, 95344, 88361]
5	Rhode Island is "almost dead last" among North...	Leonidas Raptakis	2014-02-11	7	2	[8284, 3768, 20091, 82368, 73148, 4493]
6	The poorest counties in the U.S. are in Appala...	Jim Webb	2014-11-19	8	1	[70709, 70708]
7	Koch Industries paid the legal fees of George ...		2013-07-18	9	0	[120591, 120592, 127866, 129483]
8	"Minnesota, Michigan, Iowa already have 70 mph...	Robin Vos	2013-08-22	11	1	[69547, 80095, 7994, 81116, 77621]
9	"FBI Uniform Crime Report for 2016 shows more ...	Nick Schroer	2017-10-17	12	1	[72012, 26005, 43481, 55671]
10	"Pelosi Sinks to New Low, Tells Dems: If You ...	Western Journal	2018-08-21	13	0	[27062, 27061, 20679, 61872, 20677]
11	Socialist teachers at South Charlotte Middle S...		2018-10-17	14	1	[104287, 144516]
12	Says that in the U.S. Capitol, "Stephen F. Aus...	Jonathan Saenz	2018-03-28	16	1	[16639, 16657, 16667]
13	NASA Has Just Confirmed Earth Has A New Moon	Bloggers	2018-03-29	17	0	[91455, 72179, 18903, 42080]
14	"We are always going to need architects, docto...	Mike Parson	2019-01-24	18	2	[42685, 32007, 33562]
15	"Justin Amash is rated Michigan’s No. 1 conser...	Justin Amash	2014-07-01	19	0	[22383, 72467, 72466, 86512, 73422, 83732, 83730]

	claim	claimant	date	id	label	related_articles
16	BREAKING: NFL Owner Listens to Trump, Fires P...	Multiple websites	2017-09-29	20	0	[20907, 73380, 22540, 2010]
17	Says one year ago, "no cities in the South had...	Greg Casar	2019-04-24	21	0	[87410, 18608, 57313, 35767, 85310, 43631]
18	Says North Carolina Republican Senate candidat...	Kay Hagan	2014-04-17	22	1	[81476, 67734, 73202, 96584, 73198]
19	Says "the mandate is 71 times that a child's b...	Jason Conger	2013-06-19	23	0	[87273, 87227, 11765]

In [47]:

```
# accessing the train reference articles
ARTICLES_FILEPATH = r'C:\Users\rahma\Downloads\train\train_articles'
```

In [49]:

```
df['Article'] = '\n'
df['Full_Combined'] = '\n'
```

In [50]:

```
df.head()
```

Out[50]:

	claim	claimant	date	id	label	related_articles	Article	Full_Combined
0	A line from George Orwell's novel 1984 predict...		2017-07-17	0	0	[122094, 122580, 130685, 134765]	\n	\n
1	Maine legislature candidate Leslie Gibson insu...		2018-03-17	1	2	[106868, 127320, 128060]	\n	\n
2	A 17-year-old girl named Alyssa Carson is bein...		2018-07-18	4	1	[132130, 132132, 149722]	\n	\n
3	In 1988 author Roald Dahl penned an open lette...		2019-02-04	5	2	[123254, 123418, 127464]	\n	\n
4	When it comes to fighting terrorism, "Another ...	Hillary Clinton	2016-03-22	6	2	[41099, 89899, 72543, 82644, 95344, 88361]	\n	\n

In [51]:
```
count = 0
```

In [52]:
```
cols = ['Full_Combined', 'Article']
```

In [116]:
```
# Combining the text of all reference article into a column titled 'Combined'
'''
for x in df['related_articles']:
    for i in range(len(x)):
        idx = x[i]
        with io.open(os.path.join(ARTICLES_FILEPATH, '%s.txt' % idx), 'r',encoding='cp932', errors='ignore') as f:
            df['Article'][count]=f.read()
            df['Combined'][count] += df['Article'][count]


    count +=1
'''
```

Out[116]:
```
"\n# Combining the text of all reference article into a column titled
'Combined'\nfor x in df['related_articles']:\n    for i in range(len
(x)):\n        idx = x[i]\n        with io.open(os.path.join(ARTICLES_F
ILEPATH, '%s.txt' % idx), 'r',encoding='cp932', errors='ignore') as
f:\n            df['Article'][count]=f.read()\n            df['Combine
d'][count] += df['Article'][count]\n\n    \n    count +=1 \n"
```

In [54]:
```
df.to_csv('combined.csv', encoding='utf-8')
```

In [78]:
```
df = pd.read_csv('combined.csv')

df
```

Out[78]:

Unnamed: 0	claim	claimant	date	id	label	related_articles

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
0	0	A line from George Orwell's novel 1984 predict...	NaN	2017-07-17	0	0	[122094, 122580, 130685, 134765]	
1	1	Maine legislature candidate Leslie Gibson insu...	NaN	2018-03-17	1	2	[106868, 127320, 128060]	
2	2	A 17-year-old girl named Alyssa Carson is bein...	NaN	2018-07-18	4	1	[132130, 132132, 149722]	C I
3	3	In 1988 author Roald Dahl penned an open lette...	NaN	2019-02-04	5	2	[123254, 123418, 127464]	
4	4	When it comes to fighting terrorism, "Another ...	Hillary Clinton	2016-03-22	6	2	[41099, 89899, 72543, 82644, 95344, 88361]	
5	5	Rhode Island is "almost dead last" among North...	Leonidas Raptakis	2014-02-11	7	2	[8284, 3768, 20091, 82368, 73148, 4493]	L
6	6	The poorest counties in the U.S. are in Appala...	Jim Webb	2014-11-19	8	1	[70709, 70708]	Ap
7	7	Koch Industries paid the legal fees of George ...	NaN	2013-07-18	9	0	[120591, 120592, 127866, 129483]	F
8	8	"Minnesota, Michigan, Iowa already have 70 mph...	Robin Vos	2013-08-22	11	1	[69547, 80095, 7994, 81116, 77621]	

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
9	9	"FBI Uniform Crime Report for 2016 shows more ...	Nick Schroer	2017-10-17	12	1	[72012, 26005, 43481, 55671]	M
10	10	"Pelosi Sinks to New Low, Tells Dems: If You ...	Western Journal	2018-08-21	13	0	[27062, 27061, 20679, 61872, 20677]	L
11	11	Socialist teachers at South Charlotte Middle S...	NaN	2018-10-17	14	1	[104287, 144516]	T
12	12	Says that in the U.S. Capitol, "Stephen F. Aus...	Jonathan Saenz	2018-03-28	16	1	[16639, 16657, 16667]	jb
13	13	NASA Has Just Confirmed Earth Has A New Moon	Bloggers	2018-03-29	17	0	[91455, 72179, 18903, 42080]	E
14	14	"We are always going to need architects, docto...	Mike Parson	2019-01-24	18	2	[42685, 32007, 33562]	
15	15	"Justin Amash is rated Michigan’s No. 1 conser...	Justin Amash	2014-07-01	19	0	[22383, 72467, 72466, 86512, 73422, 83732, 83730]	El F
16	16	BREAKING: NFL Owner Listens to Trump, Fires P...	Multiple websites	2017-09-29	20	0	[20907, 73380, 22540, 2010]	
17	17	Says one year ago, "no cities in the South had...	Greg Casar	2019-04-24	21	0	[87410, 18608, 57313, 35767, 85310, 43631]	
18	18	Says North Carolina Republican Senate candidat...	Kay Hagan	2014-04-17	22	1	[81476, 67734, 73202, 96584, 73198]	I

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
19	19	Says "the mandate is 71 times that a child's b...	Jason Conger	2013-06-19	23	0	[87273, 87227, 11765]	Or b
20	20	Mergers and integration in agribusiness "squee...	Elizabeth Warren	2019-03-27	24	1	[20286, 48586, 20910, 36432, 36441, 36437, 569...	
21	21	Says the Human Rights Campaign is secretly funded	Pat McCrory	2016-05-24	25	1	[19453, 48239, 48228, 69167]	
22	22	A scientific study demonstrated that conspirac...	NaN	2018-09-13	26	0	[105261, 150637, 154409]	S
23	23	Eggs and popcorn kernels can be cooked by plac...	NaN	2017-11-06	27	0	[111152, 142303]	
24	24	Says Bernie Sanders "was against the auto bail...	Hillary Clinton	2016-03-06	28	1	[95550, 56194, 60046, 78162, 21930]	r
25	25	Congress has approved the creation of a taxpay...	NaN	2016-12-29	29	1	[122509, 124298, 128806, 162047]	
26	26	In 2008, "candidate Obama, he's not even presi...	Kimberley Strassel	2017-05-28	30	0	[79629, 21682, 75456, 78190, 46593, 60479, 87559]	E
27	27	"This war has been going on for over five year...	Vitaly Churkin	2016-10-14	32	1	[115012, 122194, 115826]	
28	28	"Chicago now has City ID Cards which allow ill...	Bloggers	2019-02-28	33	0	[32175, 50380, 49922, 49555, 14758]	A

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
29	29	The Wharton School wrote an open letter to Don...	NaN	2016-07-17	34	1	[134147, 161983]	
...	...	...	...	...	...	...	...	
15525	15525	"We have an 80 percent graduation rate in high...	Jeb Bush	2015-04-17	17111	2	[79715, 4966, 32220, 92822]	
15526	15526	"Democrat Jon Ossoff would be a disaster in Co...	Donald Trump	2017-04-18	17112	1	[59229, 58243, 59052, 58923, 59238, 88895, 768...	Pr
15527	15527	Reddit postings show the shooter in Jacksonvi...	Various websites	2018-08-30	17113	0	[33298, 27066, 41639, 27071, 27072, 27070, 270...	
15528	15528	“We’re making more than ever off oil and gas ...	Jerry Patterson	2010-04-27	17114	1	[92265, 90352]	
15529	15529	"The government is trying to now close the Lin...	Glenn Beck	2010-06-28	17115	0	[78697, 86478]	
15530	15530	The Trump administration blocked public access...	NaN	2018-03-14	17116	1	[127389, 129926, 132953, 143459, 143460, 14346...	T pu
15531	15531	WalMart has put all their Christian employees ...	NaN	2015-04-08	17117	0	[108442, 114152]	Ar
15532	15532	Sen. Joe Lieberman's "home state has a public ...	Keith Olbermann	2009-10-27	17119	1	[84153, 91761]	Ol

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
15533	15533	“These are the same people that said Saddam Hu...	Donald Trump	2016-12-09	17120	1	[61264, 80263, 57526, 54087]	S s
15534	15534	"One-third of the counties — think of it, one-...	Donald Trump	2017-03-13	17121	2	[62819, 7821]	
15535	15535	Says Rep. Martha McSally "is a #FlipFlopBorder...	Kelli Ward	2018-03-16	17122	1	[67183, 20180, 41193, 28711, 20181, 34090, 201...	
15536	15536	Californians pay "the highest electricity bill...	John Cox	2017-10-21	17123	0	[72463, 32554, 82046]	C
15537	15537	Says the Steele dossier "was responsible for s...	Donald Trump	2018-07-23	17124	0	[50977, 50988, 47338]	
15538	15538	Donald Trump dropped out of the presidential r...	NaN	2016-08-20	17125	0	[91555, 91556, 91557]	
15539	15539	A photograph shows a musher riding over snowle...	NaN	2017-11-06	17126	2	[108255, 109040, 110398, 114040, 114042, 11410...	ap
15540	15540	"5.7 million -- that’s how many illegal immigr...	Ainsley Earhardt	2017-06-20	17127	0	[81362, 54803, 88380, 79869, 59084, 59582, 595...	
15541	15541	"Evidence surfaces of Vatican funding caravans...	PuppetStringNews.com	2018-11-25	17128	0	[30308, 43940]	

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
15542	15542	"The average premium across this country has ...	Mike Pence	2017-05-25	17129	0	[32457, 7958]	(
15543	15543	At an Arizona town hall event, Sen. Jeff Flake...	NaN	2017-04-17	17130	1	[143451, 145474]	
15544	15544	A photograph shows Donald Trump, Muhammad Ali,...	NaN	2018-08-13	17131	2	[118938, 125644, 127592]	
15545	15545	A photograph shows a man mowing his lawn durin...	NaN	2017-06-05	17132	2	[107244, 115705, 142189]	
15546	15546	President Obama signed a law permanently prote...	NaN	2017-01-09	17133	1	[107369, 122972, 147969, 38987, 151939]	
15547	15547	"I haven’t really proposed (phasing out aid to...	Rand Paul	2014-08-04	17134	0	[88399, 91476, 11371, 91483, 7021]	Ra
15548	15548	Says Aaron Rodgers "is not the highest tax rat...	Paul Ryan	2017-08-21	17135	1	[53671, 30934, 94982, 30953, 30949]	D
15549	15549	"They (Clinton and Obama) have never to my kno...	John McCain	2008-05-13	17136	0	[67611, 67699, 67610, 82239, 86166, 3653, 7112...	(
15550	15550	The omnibus spending bill has "9,427 pork barr...	John McCain	2009-02-25	17137	2	[82947, 93503]	

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
15551	15551	Representative Maxine Waters said Muslims were...	NaN	2017-06-06	17138	0	[103780, 104726, 126025]	T Br
15552	15552	"We were not, I repeat, were not told that wat...	Nancy Pelosi	2009-04-23	17139	0	[11331, 68915, 2186, 2185, 88418, 81950]	su
15553	15553	As of August 2017, members of the public could...	NaN	2018-05-14	17140	2	[121353, 152864, 154411]	li
15554	15554	"We don't get any of that information" from th...	Scott Walker	2016-12-23	17141	1	[69545, 88929, 14698]	

15555 rows × 8 columns

In [79]:

```
# Claimant will be used as a feature later. Hence, all entries with NaN
 claimants are are being deleted.
df = df[pd.notnull(df['claimant'])]
```

In [80]:

```
df.shape
```

Out[80]:

```
(10593, 8)
```

In [81]:

```
df = df[pd.notnull(df['Combined'])]
df.shape
#No NaN in Combined column
```

Out[81]:

```
(10593, 8)
```

In [82]:

```
df
```

Out[82]:

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
4	4	When it comes to fighting terrorism, "Another ...	Hillary Clinton	2016-03-22	6	2	[41099, 89899, 72543, 82644, 95344, 88361]	
5	5	Rhode Island is "almost dead last" among North...	Leonidas Raptakis	2014-02-11	7	2	[8284, 3768, 20091, 82368, 73148, 4493]	
6	6	The poorest counties in the U.S. are in Appala...	Jim Webb	2014-11-19	8	1	[70709, 70708]	
8	8	"Minnesota, Michigan, Iowa already have 70 mph...	Robin Vos	2013-08-22	11	1	[69547, 80095, 7994, 81116, 77621]	I
9	9	"FBI Uniform Crime Report for 2016 shows more ...	Nick Schroer	2017-10-17	12	1	[72012, 26005, 43481, 55671]	
10	10	"Pelosi Sinks to New Low, Tells Dems: If You ...	Western Journal	2018-08-21	13	0	[27062, 27061, 20679, 61872, 20677]	I
12	12	Says that in the U.S. Capitol, "Stephen F. Aus...	Jonathan Saenz	2018-03-28	16	1	[16639, 16657, 16667]	
13	13	NASA Has Just Confirmed Earth Has A New Moon	Bloggers	2018-03-29	17	0	[91455, 72179, 18903, 42080]	
14	14	"We are always going to need architects, docto...	Mike Parson	2019-01-24	18	2	[42685, 32007, 33562]	S

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
15	15	"Justin Amash is rated Michigan's No. 1 conser...	Justin Amash	2014-07-01	19	0	[22383, 72467, 72466, 86512, 73422, 83732, 83730]	
16	16	BREAKING: NFL Owner Listens to Trump, Fires P...	Multiple websites	2017-09-29	20	0	[20907, 73380, 22540, 2010]	
17	17	Says one year ago, "no cities in the South had...	Greg Casar	2019-04-24	21	0	[87410, 18608, 57313, 35767, 85310, 43631]	La
18	18	Says North Carolina Republican Senate candidat...	Kay Hagan	2014-04-17	22	1	[81476, 67734, 73202, 96584, 73198]	F \
19	19	Says "the mandate is 71 times that a child's b...	Jason Conger	2013-06-19	23	0	[87273, 87227, 11765]	
20	20	Mergers and integration in agribusiness "squee...	Elizabeth Warren	2019-03-27	24	1	[20286, 48586, 20910, 36432, 36441, 36437, 569...	L
21	21	Says the Human Rights Campaign is secretly funded	Pat McCrory	2016-05-24	25	1	[19453, 48239, 48228, 69167]	E 2
24	24	Says Bernie Sanders "was against the auto bail...	Hillary Clinton	2016-03-06	28	1	[95550, 56194, 60046, 78162, 21930]	re
26	26	In 2008, "candidate Obama, he's not even presi...	Kimberley Strassel	2017-05-28	30	0	[79629, 21682, 75456, 78190, 46593, 60479, 87559]	

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
27	27	"This war has been going on for over five year...	Vitaly Churkin	2016-10-14	32	1	[115012, 122194, 115826]	
28	28	"Chicago now has City ID Cards which allow ill...	Bloggers	2019-02-28	33	0	[32175, 50380, 49922, 49555, 14758]	C
30	30	"If you're from Syria and you're a Christian, ...	Donald Trump	2015-07-11	35	0	[2248, 91899, 94849, 79994, 32228]	D
31	31	"Expanding Medicaid would require borrowing mo...	Will Weatherford	2013-05-09	36	1	[66749, 1228, 7897, 10786]	ne
32	32	U.S. Reps. John Barrow and Sanford Bishop and ...	National Republican Congressional Committee	2011-05-18	37	0	[12142, 12143]	
33	33	President Barack Obama "ordered our military t...	Allen West	2014-09-26	38	1	[77811, 93931, 89880, 71518, 95458]	D
34	34	The GOP’s Obamacare replacement would reduce s...	Ron Johnson	2017-03-21	39	2	[81259, 7818, 7880, 28679, 47758]	J
37	37	Saturday Night Live executive producer "Lorne ...	Ted Cruz	2014-09-09	42	1	[67429, 70724, 87380, 82113]	Fa
38	38	Van Jones signed a petition indicating he "thi...	Glenn Beck	2009-09-03	43	1	[69944, 2170, 2171, 88297, 8595, 77947, 76149]	
39	39	Republicans approved 12 times larger tax break...	Peter Barca	2014-11-07	44	1	[69546, 81156]	El

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
40	40	“Why do we seem to have vocal and proactive op...	Vladimir Putin	2017-12-15	45	1	[129437, 149764, 136978, 136976]	R
42	42	"Only three in 10 young Americans under 30 -- ...	Ron Meyer	2013-04-29	47	1	[85160, 69061, 87988]	N
...	...	...	...	...	...	...	...	
15512	15512	It has been reported that South Africa spends ...	ThisDay Newspaper, Nigeria	2018-04-05	17098	0	[160662, 150331, 134822, 129912, 129996, 13010...	
15513	15513	The United States had "allies lined up" for ai...	Peter King	2014-08-31	17099	1	[68585, 1893, 93826, 66585, 71224, 90065, 7972...	
15514	15514	"The House of Representatives just voted 300-1...	Facebook posts	2015-06-29	17100	2	[70586, 80398, 63421, 33936, 90281, 43386, 568...	S
15516	15516	California’s Capitol building is "second only ...	Richard Pan	2016-04-07	17102	0	[45240, 87472, 92278, 2837, 86966]	(
15517	15517	"Wild Bill Hickok had his first duel in the to...	Barack Obama	2008-07-30	17103	2	[86635, 85360]	F
15518	15518	“Hillary Clinton six months ago said the vets...	Donald Trump	2016-09-07	17104	0	[85032, 37932, 76665, 76768, 38014, 6307, 3788...	
15519	15519	"The sanctions that we put on (Russia) for the...	Anthony Tata	2017-02-19	17105	1	[41721, 46676, 61120, 72216]	

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
15521	15521	"EPA officials have commended (Koch Industries...	Charles Koch	2014-04-02	17107	1	[96957, 76517, 96955, 96964, 69774]	F
15523	15523	"Santorum also voted for a teapot museum in No...	Rick Perry	2012-01-02	17109	1	[54296, 2949]	U
15524	15524	Says President Dwight Eisenhower "moved 1.5 mi...	Donald Trump	2015-11-10	17110	1	[23570, 59621, 35794]	As
15525	15525	"We have an 80 percent graduation rate in high...	Jeb Bush	2015-04-17	17111	2	[79715, 4966, 32220, 92822]	
15526	15526	"Democrat Jon Ossoff would be a disaster in Co...	Donald Trump	2017-04-18	17112	1	[59229, 58243, 59052, 58923, 59238, 88895, 768...	
15527	15527	Reddit postings show the shooter in Jacksonvi...	Various websites	2018-08-30	17113	0	[33298, 27066, 41639, 27071, 27072, 27070, 270...	
15528	15528	“We’re making more than ever off oil and gas ...	Jerry Patterson	2010-04-27	17114	1	[92265, 90352]	
15529	15529	"The government is trying to now close the Lin...	Glenn Beck	2010-06-28	17115	0	[78697, 86478]	8
15532	15532	Sen. Joe Lieberman's "home state has a public ...	Keith Olbermann	2009-10-27	17119	1	[84153, 91761]	

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
15533	15533	“These are the same people that said Saddam Hu...	Donald Trump	2016-12-09	17120	1	[61264, 80263, 57526, 54087]	s
15534	15534	"One-third of the counties — think of it, one-...	Donald Trump	2017-03-13	17121	2	[62819, 7821]	
15535	15535	Says Rep. Martha McSally "is a #FlipFlopBorder...	Kelli Ward	2018-03-16	17122	1	[67183, 20180, 41193, 28711, 20181, 34090, 201...	
15536	15536	Californians pay "the highest electricity bill...	John Cox	2017-10-21	17123	0	[72463, 32554, 82046]	
15537	15537	Says the Steele dossier "was responsible for s...	Donald Trump	2018-07-23	17124	0	[50977, 50988, 47338]	
15540	15540	"5.7 million -- that’s how many illegal immigr...	Ainsley Earhardt	2017-06-20	17127	0	[81362, 54803, 88380, 79869, 59084, 59582, 595...	
15541	15541	"Evidence surfaces of Vatican funding caravans...	PuppetStringNews.com	2018-11-25	17128	0	[30308, 43940]	C
15542	15542	"The average premium across this country has ...	Mike Pence	2017-05-25	17129	0	[32457, 7958]	
15547	15547	"I haven’t really proposed (phasing out aid to...	Rand Paul	2014-08-04	17134	0	[88399, 91476, 11371, 91483, 7021]	

	Unnamed: 0	claim	claimant	date	id	label	related_articles	
15548	15548	Says Aaron Rodgers "is not the highest tax rat...	Paul Ryan	2017-08-21	17135	1	[53671, 30934, 94982, 30953, 30949]	
15549	15549	"They (Clinton and Obama) have never to my kno...	John McCain	2008-05-13	17136	0	[67611, 67699, 67610, 82239, 86166, 3653, 7112...	(
15550	15550	The omnibus spending bill has "9,427 pork barr...	John McCain	2009-02-25	17137	2	[82947, 93503]	s
15552	15552	"We were not, I repeat, were not told that wat...	Nancy Pelosi	2009-04-23	17139	0	[11331, 68915, 2186, 2185, 88418, 81950]	Pe
15554	15554	"We don't get any of that information" from th...	Scott Walker	2016-12-23	17141	1	[69545, 88929, 14698]	

10593 rows × 8 columns

Data Cleaning and Feature Selection/Engineering

In [83]:

```
# Data cleaning and pre-process dataset
nltk.download('stopwords')

# TEXT CLENAING
TEXT_CLEANING_RE = "\(|\)""!@#$%^&*<>?/.,;:|=@\S+|https?:\S+|http?:\S|
[^A-Za-z0-9]+"
stop_words = stopwords.words("english")
stemmer = SnowballStemmer("english")
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\rahma\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
[nltk_data]   Package stopwords is already up-to-date!
```

In [84]:

```
# Applying data cleaning across claim and combined article columns
def preprocess(text, stem=False):
    # Remove link,user and special characters
    text = re.sub(TEXT_CLEANING_RE, ' ', str(text).lower()).strip()
    tokens = []
    for token in text.split():
        if token not in stop_words:
            if stem:
                tokens.append(stemmer.stem(token))
            else:
                tokens.append(token)
    return " ".join(tokens)

df.claim = df.claim.apply(lambda x: preprocess(x))
df.Combined = df.Combined.apply(lambda x: preprocess(x))
```

In [131]:

```
df.claim
```

Out[131]:

```
4        comes fighting terrorism another thing know wo...
5        rhode island almost dead last among northeaste...
6        poorest counties u appalachia happen 90 percen...
8        minnesota michigan iowa already 70 mph speed l...
9        fbi uniform crime report 2016 shows four times...
10          pelosi sinks new low tells dems lie voters win
12       says u capitol stephen f austin sam houston st...
13                           nasa confirmed earth new moon
14       always going need architects doctors going nee...
15       justin amash rated michigan 1 conservative nat...
16       breaking nfl owner listens trump fires player ...
17       says one year ago cities south guaranteed paid...
18       says north carolina republican senate candidat...
19       says mandate 71 times child body injected dise...
20       mergers integration agribusiness squeeze famil...
21              says human rights campaign secretly funded
24       says bernie sanders auto bailout voted money e...
26       2008 candidate obama even president elect send...
27       war going five years destruction see aleppo ea...
```

```
28       chicago city id cards allow illegal immigrants...
30             syria christian cannot come country refugee
31       expanding medicaid would require borrowing mon...
32       u reps john barrow sanford bishop fellow democ...
33       president barack obama ordered military enlist...
34       gop obamacare replacement would reduce subsidi...
37       saturday night live executive producer lorne m...
38       van jones signed petition indicating thinks bu...
39       republicans approved 12 times larger tax break...
40       seem vocal proactive opposition members countr...
42       three 10 young americans 30 30 percent 30 full...
                               ...
15512    reported south africa spends seven times per h...
15513    united states allies lined air strikes syria o...
15514    house representatives voted 300 131 remove cou...
15516    california capitol building second disneyland ...
15517    wild bill hickok first duel town square family...
15518    hillary clinton six months ago said vets treat...
15519    sanctions put russia crimea annexation meddlin...
15521    epa officials commended koch industries commit...
15523     santorum also voted teapot museum north carolina
15524    says president dwight eisenhower moved 1 5 mil...
15525    80 percent graduation rate high school spendin...
15526    democrat jon ossoff would disaster congress we...
15527    reddit postings show shooter jacksonville flor...
15528    making ever oil gas right secret oil productio...
15529    government trying close lincoln memorial kind ...
15532    sen joe lieberman home state public option cov...
15533    people said saddam hussein weapons mass destru...
15534    one third counties think one third one insurer...
15535           says rep martha mcsally flipflopborderhawk
15536    californians pay highest electricity bills nation
15537    says steele dossier responsible starting speci...
15540    5 7 million many illegal immigrants might vote...
15541    evidence surfaces vatican funding caravans tar...
15542    average premium across country actually double...
15547              really proposed phasing aid israel past
15548    says aaron rodgers highest tax rate payer wisc...
15549    clinton obama never knowledge involved legisla...
```

```
15550        omnibus spending bill 9 427 pork barrel items
15552    repeat told waterboarding enhanced interrogati...
15554    get information federal government refugees co...
Name: claim, Length: 10593, dtype: object
```

In [133]:
```
df.Combined
```

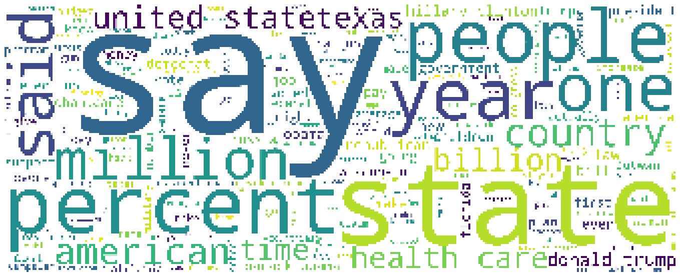
Out[133]:
```
4        remarks counterterrorism stanford university l...
5        lis code virginia 18 2 10 prev next 18 2 10 pu...
6        counties appalachia alabama bibb blount calhou...
8        robin vos discusses milwaukee crime speed limi...
9        fbi four times people stabbed death killed rif...
10       pelosi sinks new low tells dems lie voters win...
12       0418 jblancpftexas emails jennifer blancato cu...
13       another moon earth well really depends point v...
14       seg 1 missouri governor seeks efficiency seg 2...
15       elections election type president united state...
16       statement nfl commissioner roger goodell page ...
17       current sick days laws paid sick days laws soo...
18       radio ad tillis may values e north carolina ka...
19       oregon house approves bill tightening rules pa...
20       leveling playing field america family farmers ...
21       equality magazine spring 2016 h u n r g h c p ...
24       vpr leahy sanders reluctantly support auto ind...
26       life secret back channel iran secret back chan...
27       russian arms shipments bolster syria embattled...
28       citykey chicago citykey optional valid governm...
30       donald trump tells brody file president greate...
31       responsible safety net florida legislature con...
32       blue dog coalition contact 202 226 9782 blue d...
33       dod planning let illegal immigrants enlist def...
34       wisconsin sen ron johnson skeptical house gop ...
37       fact checking sen cruz claim potential ban snl...
38       911 truth statement respected leaders families...
39       peter barca reacts election results episode tr...
40       russia economic report summary real gdp growth...
42       ncoc illennials play central role nation civic...
                               ...
15512    nigeria poor attitude healthcare financing man...
```

```
15513    former vice president dick cheney speaks face ...
15514    agricultural marketing service version 1 0 enc...
15516    yosemite national park u national park service...
15517    mouth potomac barack obama mocked hillary clin...
15518    nbc news presents first ever commander chief f...
15519    rub usd historical data rub usd russian ruble ...
15521    10 22 2010 flint hills resources lp agrees tra...
15523    u senate u senate roll call votes 109th congre...
15524    texas state historical association tsha fred l...
15525    study us education spending tops global list s...
15526    president trump king flip flops continued fact...
15527    jacksonville shooter history mental illness po...
15528    patterson future green energy nearly four year...
15529    glenn attacked 8 28 glenn beck seen glennbeck ...
15532    countdown keith olbermann tuesday october 27 2...
15533    secret cia assessment says russia trying help ...
15534    remarks president trump listening session heal...
15535    border district republicans skeptical trump wa...
15536    cox pins gubernatorial campaign eighborhood le...
15537    democrats memo pushes back gop claims russia p...
15540    immigration facts general government data crim...
15541    new funding helps catholic ministries provide ...
15542    centers medicare medicaid services health insu...
15547    rand paul end welfare israel jennifer epstein ...
15548    dor tax rates sales tax rate chart following c...
15549    1389 110th congress 2007 2008 climate change e...
15550    earmark reform 2009 spending bill contains 9 0...
15552    pelosi disputes suggestion told waterboarding ...
15554    gov scott walker previews next year state gove...
Name: Combined, Length: 10593, dtype: object
```

In [85]:

```
# all claims
all_words = " ".join(df.claim)

# Wordcloud of claims
wordcloud = WordCloud(height=4000, width=10000, stopwords=STOPWORDS, ba
ckground_color='white')
wordcloud = wordcloud.generate(all_words)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
```

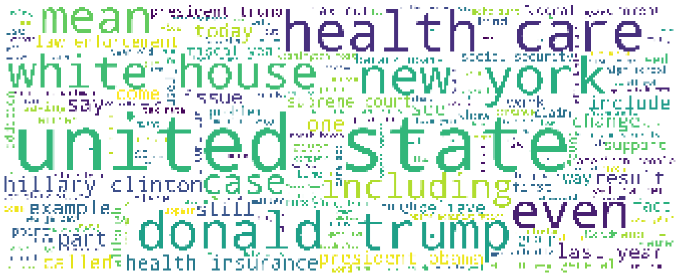


As we can see from the above, most of the claims state some sort of fact often referencing other sources, as words such "say", "year", "percent", "million" are some of the more common ones.

In [86]:

```
# all reference articles
all_words = " ".join(df.Combined)

# Wordcloud of combined articles
wordcloud = WordCloud(height=4000, width=10000, stopwords=STOPWORDS, ba
ckground_color='white')
wordcloud = wordcloud.generate(all_words)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
```



Compared to the claim column, there is much less dominance of single words in the combined reference article column. Some expected common words include United States itself, major issues such as "Health Care" and political personnel as "Donald Trump".

```
In [87]:  # One hot label encoding for claimant.
          claimant=df['claimant']
          claimant_encoded=pd.get_dummies(claimant)
          claimant_encoded.shape

Out[87]:  (10593, 3104)
```

```
In [136]: claimant_encoded.reset_index(drop=True, inplace=True)
          claimant_encoded.shape

Out[136]: (10593, 3104)
```

```
In [89]:  # Converting dates to quarters to reduce future feature size
          df['Qtr'] = pd.to_datetime(df['date'].values, format='%Y-%m').astype('p
          eriod[Q]')
```

```
In [135]: # One hot label encoding for date (Quarter)
          dfqtr_encoded=pd.get_dummies(df['Qtr'])
          dfqtr_encoded.reset_index(drop=True, inplace=True)
          dfqtr_encoded.shape

Out[135]: (10593, 52)
```

```
In [91]:  # Creating a column which calculates the number of articles referenced
           for each entry
          df['article_no']=df.related_articles.astype(str).str.count(",")+1
```

```
In [134]: # One hot encoding of article_no
          dfart_encoded=pd.get_dummies(df['article_no'])
          dfart_encoded.reset_index(drop=True, inplace=True)
          dfart_encoded.shape

Out[134]: (10593, 39)
```

Data Exploration/Visualization

In [137]:

```
# Dataframe of all false news
df_0=df.loc[df['label'] == 0]
df_0.shape
```

Out[137]:

```
(4374, 10)
```

In [138]:

```
#Dataframe of all partly true news
df_1=df.loc[df['label'] == 1]
df_1.shape
```

Out[138]:

```
(5164, 10)
```

In [139]:

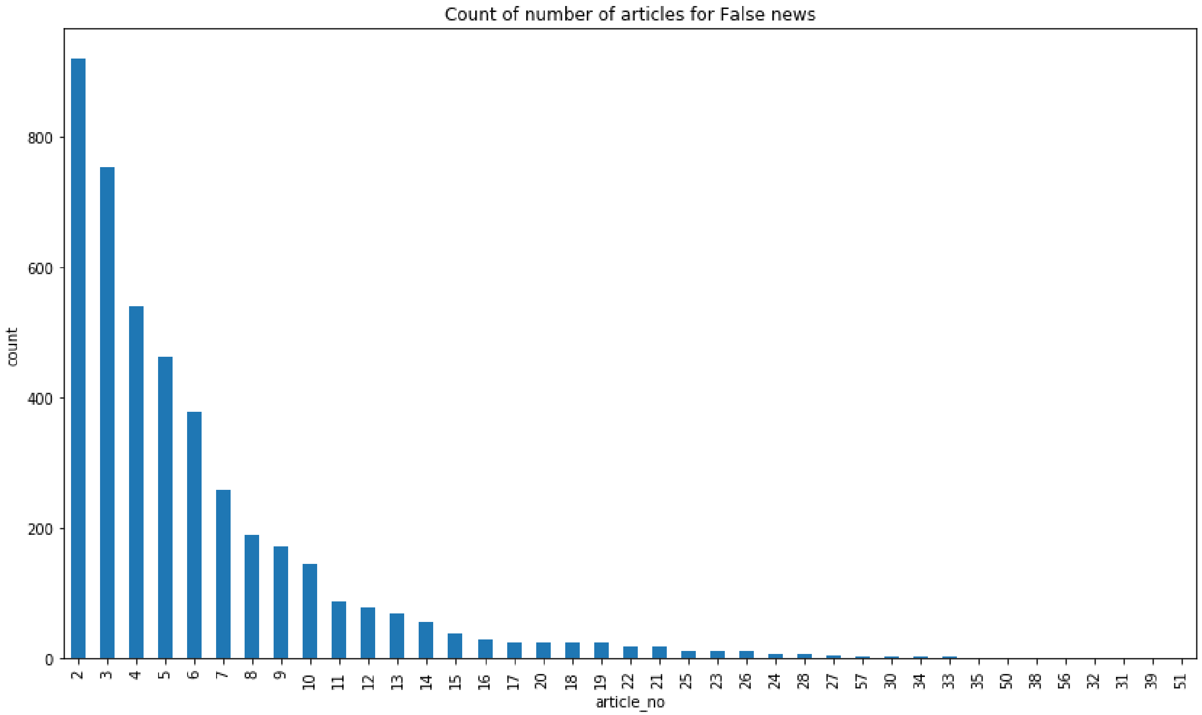
```
#Dataframe of all true news
df_2=df.loc[df['label'] == 2]
df_2.shape
```

Out[139]:

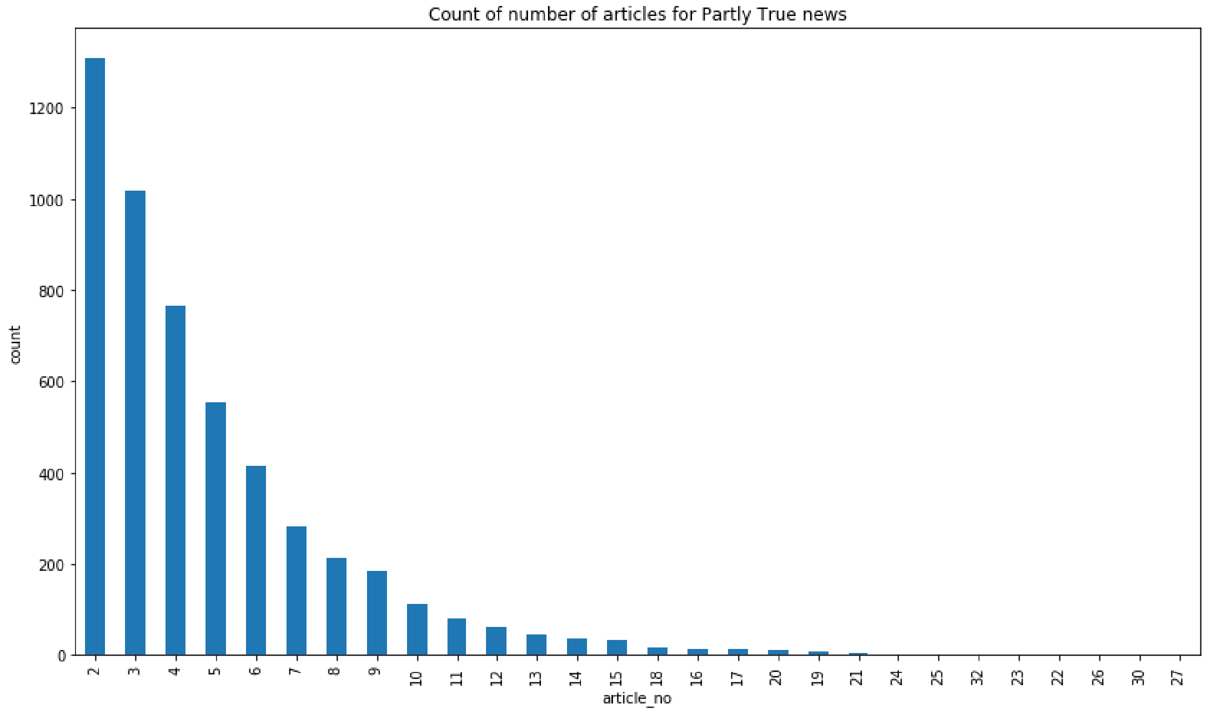
```
(1055, 10)
```

In [96]:

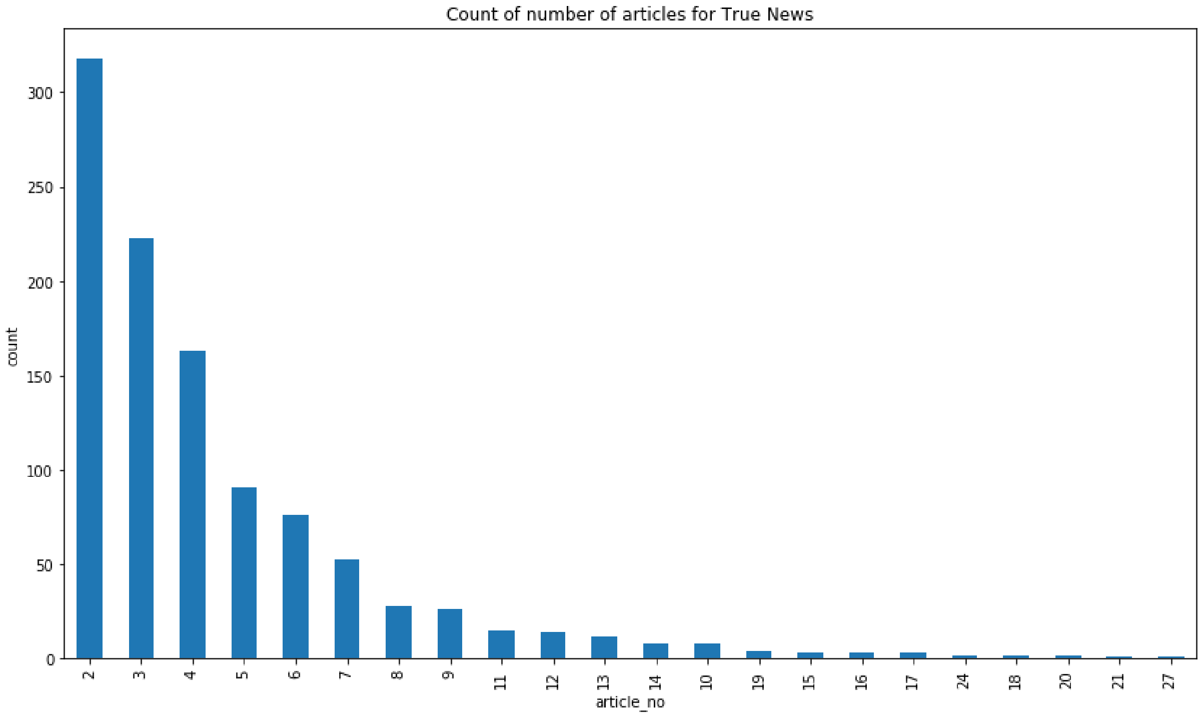
```
ax = df_0['article_no'].value_counts().plot(kind='bar',
                                    figsize=(14,8),
                                    title="Count of number of articles
 for False news")
ax.set_xlabel("article_no")
ax.set_ylabel("count")
plt.show()
```



In [97]:
```
ax = df_1['article_no'].value_counts().plot(kind='bar',
                                              figsize=(14,8),
                                              title="Count of number of articles
 for Partly True news")
ax.set_xlabel("article_no")
ax.set_ylabel("count")
plt.show()
```



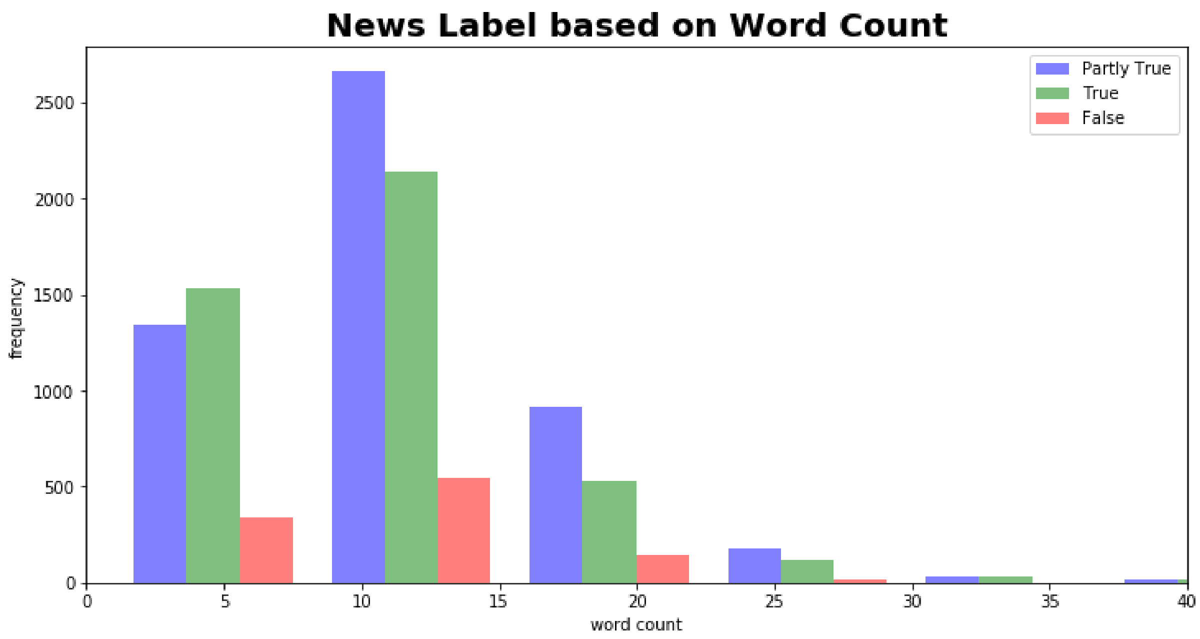
In [98]:
```
ax = df_2['article_no'].value_counts().plot(kind='bar',
                                                   figsize=(14,8),
                                                   title="Count of number of articles
 for True News")
ax.set_xlabel("article_no")
ax.set_ylabel("count")
plt.show()
```



In [149]:

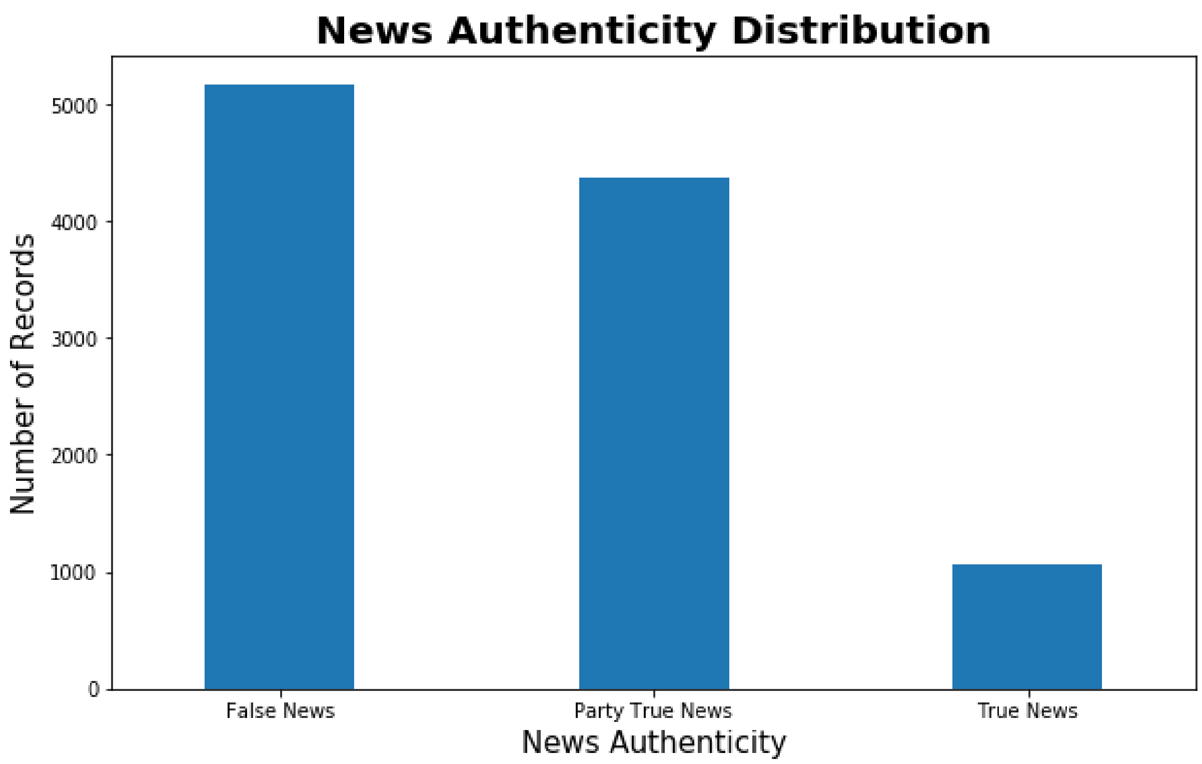
```
def word_count(sentence):
    return len(sentence.split())
df['word count'] = df['claim'].apply(word_count)
#plot word count distribution for both positive and negative sentiments
x = df['word count'][df['label'] == 1]
y = df['word count'][df['label'] == 0]
z = df['word count'][df['label'] == 2]
plt.figure(figsize=(12,6))
plt.xlim(0,40)
plt.xlabel('word count')
plt.ylabel('frequency')
plt.title('News Label based on Word Count',fontsize=20,weight='bold')
g = plt.hist([x, y, z], color=['b','g','r'], alpha=0.5, label=['Partly True','True', 'False'])
plt.legend(loc='upper right')
```

```
Out[149]: <matplotlib.legend.Legend at 0x296eeefb358>
```



```
In [157]: #Bar plot to visualize the distribution of cleaned data
fig_1=df['label'].value_counts().plot(kind='bar',width=0.4,figsize=(10,
6))
plt.xlabel('News Authenticity',fontsize=15)
plt.ylabel('Number of Records',fontsize=15)
plt.title('News Authenticity Distribution',fontsize=20,weight='bold')
labels = [item.get_text() for item in fig_1.get_xticklabels()]
labels[0] = 'False News'
labels[1] = 'Party True News'
labels[2] = 'True News'
fig_1.set_xticklabels(labels,rotation='horizontal')
```

```
Out[157]: [Text(0, 0, 'False News'),
           Text(0, 0, 'Party True News'),
           Text(0, 0, 'True News')]
```

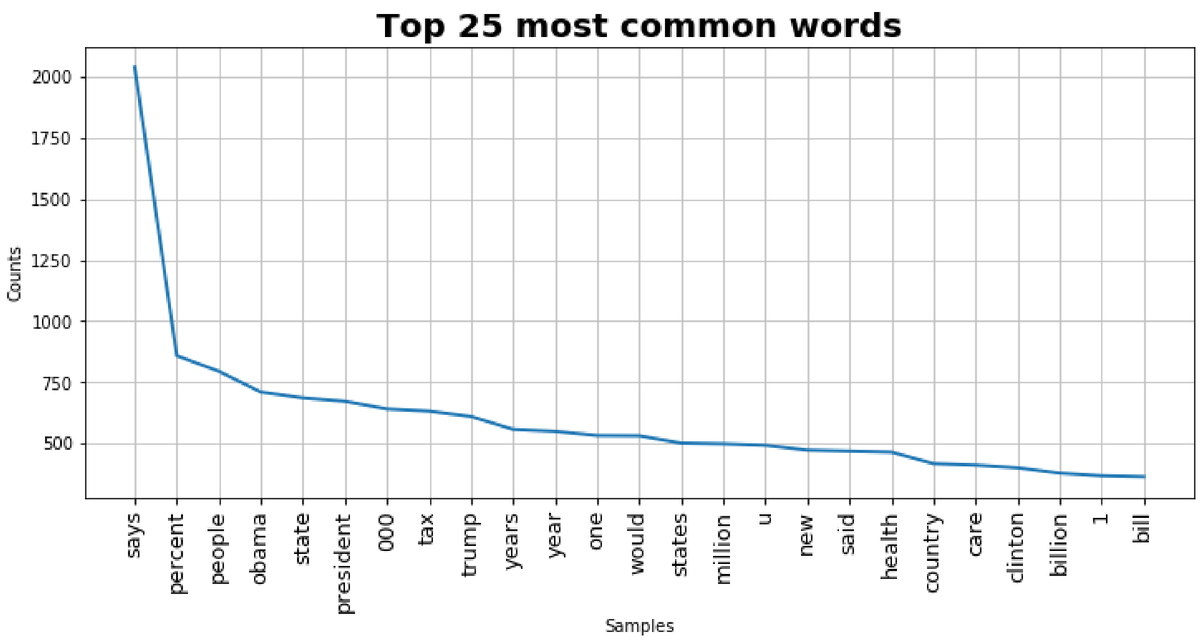


In [156]:

```
#Get most common words in training dataset
all_words = []
for line in list(df['claim']):
    words = line.split()
    for word in words:
        all_words.append(word.lower())
        #Split sentences to get individual words

# Create a word frequency dictionary
wordfreq = Counter(all_words)

#Plot word frequency distribution of first few words
plt.figure(figsize=(12,5))
plt.title('Top 25 most common words',fontsize=20,weight='bold')
plt.xticks(fontsize=13, rotation=90)
fd = nltk.FreqDist(all_words)
fd.plot(25,cumulative=False)
```



```
Out[156]: <matplotlib.axes._subplots.AxesSubplot at 0x296eea9cd30>
```

Model Implementation

```
In [99]: # tokenize all the cleaned claims in our dataset. Tokens are individual
          terms or words,
         # and tokenization is the process of splitting a string of text into to
         kens
         tokenized_tweet = df['claim'].apply(lambda x: x.split())
         tokenized_tweet.head()
```

```
Out[99]: 4    [comes, fighting, terrorism, another, thing, k...
         5    [rhode, island, almost, dead, last, among, nor...
         6    [poorest, counties, u, appalachia, happen, 90,...
         8    [minnesota, michigan, iowa, already, 70, mph, ...
```

```
9    [fbi, uniform, crime, report, 2016, shows, fou...
Name: claim, dtype: object
```

In [100]:

```
# Stemming is a rule-based process of stripping the suffixes (“ing”, “l
y”, “es”, “s” etc) from a word
from nltk.stem.porter import *
stemmer = PorterStemmer()

tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for
i in x]) # stemming
tokenized_tweet.head()
```

Out[100]:

```
4    [come, fight, terror, anoth, thing, know, work...
5    [rhode, island, almost, dead, last, among, nor...
6    [poorest, counti, u, appalachia, happen, 90, p...
8    [minnesota, michigan, iowa, alreadi, 70, mph, ...
9    [fbi, uniform, crime, report, 2016, show, four...
Name: claim, dtype: object
```

In [101]:

```
# TF-IDF works by penalizing the common words by assigning them lower w
eights while giving
# importance to words which are rare in the entire corpus but appear in
 good numbers
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vectorizer = TfidfVectorizer(max_df=0.90, min_df=2, max_features=
1000, stop_words='english')
# TF-IDF feature matrix
tfidf = tfidf_vectorizer.fit_transform(df['claim'])
print(tfidf)
```

```
  (0, 234)	0.39288101322624264
  (0, 385)	0.38433898683811474
  (0, 989)	0.3580476166910551
  (0, 994)	0.35111189074335103
  (0, 544)	0.2857149928057158
  (0, 1084)	0.3041601634454739
  (0, 127)	0.3719738249990458
  (0, 365)	0.36597733314713427
```

```
(0, 365)      0.36597733314713427
(1, 520)      0.49770138111258677
(1, 297)      0.45940247477041085
(1, 947)      0.27991761613023297
(1, 1003)     0.31421800797831284
(1, 934)      0.4236649514889044
(1, 763)      0.42972535640273585
(2, 264)      0.5254603539389164
(2, 448)      0.5285033147463913
(2, 53)       0.4640245127728134
(2, 722)      0.2771771024423438
(2, 1076)     0.3904225507934768
(3, 643)      0.5334923838254498
(3, 633)      0.5334923838254498
(3, 48)       0.46143333842637607
(3, 56)       0.4667451413664284
(4, 381)      0.3269605749451925
(4, 278)      0.31704124153757257
:      :
(10588, 590)  0.2648086968032802
(10588, 51)   0.25421387790366945
(10588, 424)  0.27033646717335785
(10588, 973)  0.29018925093644515
(10588, 131)  0.256267031819675
(10588, 31)   0.2971047529862
(10588, 192)  0.2815936487658725
(10589, 684)  0.24263719453502747
(10589, 225)  0.4102014751194491
(10589, 195)  0.3761081368836556
(10589, 226)  0.2844755484417136
(10589, 522)  0.43146083069161917
(10589, 514)  0.4631783254768138
(10589, 566)  0.38702152877337914
(10590, 935)  1.0
(10591, 1036) 0.6683231907972086
(10591, 1006) 0.7438710322647588
(10592, 382)  0.27570176304226185
(10592, 233)  0.35459492616133687
(10592, 505)  0.36927137206614075
(10592, 431)  0.26823907096449745
```

```
  (10592, 817)	0.3724786818230002

  (10592, 587)	0.3605747395757338
  (10592, 949)	0.4231083645931284
  (10592, 578)	0.3770440708393888
```

In [102]:

```
# tokenize all the cleaned Combined articles in our dataset. Tokens are
 individual terms or words,
# and tokenization is the process of splitting a string of text into to
kens
tokenized_tweet = df['Combined'].apply(lambda x: x.split())
tokenized_tweet.head()
```

Out[102]:

```
4    [remarks, counterterrorism, stanford, universi...
5    [lis, code, virginia, 18, 2, 10, prev, next, 1...
6    [counties, appalachia, alabama, bibb, blount, ...
8    [robin, vos, discusses, milwaukee, crime, spee...
9    [fbi, four, times, people, stabbed, death, kil...
Name: Combined, dtype: object
```

In [103]:

```
# Stemming is a rule-based process of stripping the suffixes (“ing”, “l
y”, “es”, “s” etc) from a word
from nltk.stem.porter import *
stemmer = PorterStemmer()

tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for
i in x]) # stemming
tokenized_tweet.head()
```

Out[103]:

```
4    [remark, counterterror, stanford, univers, log...
5    [li, code, virginia, 18, 2, 10, prev, next, 18...
6    [counti, appalachia, alabama, bibb, blount, ca...
8    [robin, vo, discuss, milwauke, crime, speed, l...
9    [fbi, four, time, peopl, stab, death, kill, ri...
Name: Combined, dtype: object
```

In [104]:

```
# TF-IDF works by penalizing the common words by assigning them lower w
eights while giving
```

```
# importance to words which are rare in the entire corpus but appear in
 good numbers
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vectorizer = TfidfVectorizer(max_df=0.90, min_df=2, max_features=
1800, stop_words='english')
# TF-IDF feature matrix
tfidf2 = tfidf_vectorizer.fit_transform(df['Combined'])
print(tfidf2)
```

```
  (0, 1543)	0.015526496027887885
  (0, 1896)	0.04327310870637819
  (0, 1760)	0.029726454606675036
  (0, 200)	0.10441308651317295
  (0, 1225)	0.016728398286438313
  (0, 1762)	0.034632221946958
  (0, 1055)	0.05910365177647567
  (0, 1937)	0.017238253887138403
  (0, 1970)	0.020638447370608156
  (0, 1824)	0.15045964527742328
  (0, 249)	0.07496013317944802
  (0, 888)	0.007594822576212721
  (0, 406)	0.013373665825399034
  (0, 1841)	0.07185955294139641
  (0, 1736)	0.026556455028765295
  (0, 689)	0.1151857507364533
  (0, 190)	0.09248356271057492
  (0, 1575)	0.007068864180469558
  (0, 986)	0.1971353047807111
  (0, 1725)	0.0067945519588095285
  (0, 1829)	0.049504992370952454
  (0, 1173)	0.047825732670316076
  (0, 1668)	0.020489009026560173
  (0, 507)	0.05888540966082406
  (0, 1314)	0.00897742047057181
  :	:
  (10592, 1765)	0.01360819502777221
  (10592, 663)	0.012722655510010528
  (10592, 521)	0.07194920648592537
  (10592, 617)	0.015895821920450033
  (10592, 667)	0.018914820785869296
  (10592, 730)	0.019335531075670816
  (10592, 1409)	0.013396387544581 76
  (10592, 977)	0.013617198836375199
  (10592, 756)	0.01375635481872194
  (10592, 1731)	0.01941208023146298
  (10592, 1864)	0.016928711444218803
  (10592, 1515)	0.20588684358553352
  (10592, 1023)	0.013937162808910736
  (10592, 1384)	0.014638625903698415
  (10592, 1711)	0.02862485833869352
  (10592, 1792)	0.04049931278553717
  (10592, 450)	0.05153617858134226
  (10592, 631)	0.016457404008293975
  (10592, 927)	0.01747554678571258
  (10592, 394)	0.017516430170172653
  (10592, 587)	0.0432898279825252
  (10592, 1284)	0.030022601557499754
  (10592, 1672)	0.08939572224777644
  (10592, 1979)	0.016480039462317228
  (10592, 1588)	0.03500553753577546
```

In [140]:

```
tfidf = tfidf.todense()
tfidf = pd.DataFrame(tfidf)
tfidf.reset_index(drop=True, inplace=True)
tfidf
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call l
ast)
<ipython-input-140-807fd83b01d6> in <module>
----> 1 tfidf = tfidf.todense()
      2 tfidf = pd.DataFrame(tfidf)
      3 tfidf.reset_index(drop=True, inplace=True)
      4 tfidf.shape

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\generic.py in __
getattr__(self, name)
   5065             if self._info_axis._can_hold_identifiers_and_holds_
```

```
name(name):
   5066                 return self[name]
-> 5067             return object.__getattribute__(self, name)
   5068 
   5069     def __setattr__(self, name, value):

AttributeError: 'DataFrame' object has no attribute 'todense'
```

In [106]:

```
tfidf2 = tfidf2.todense()
tfidf2 = pd.DataFrame(tfidf2)
tfidf2.reset_index(drop=True, inplace=True)
tfidf2
```

Out[106]:

	0	1	2	3	4	5	6	7	8
0	0.000000	0.012138	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.006720	0.007857	0.005639	0.006723	0.004996	0.002531	0.003228	0.005048	0.000000
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.000000	0.015299	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.011248	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	0.000000	0.074757	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	0.000000	0.000000	0.023895	0.024928	0.000000	0.075069	0.071829	0.000000	0.037801
10	0.000000	0.014475	0.000000	0.000000	0.000000	0.000000	0.017843	0.000000	0.000000
11	0.000000	0.015387	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
12	0.000000	0.012081	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
13	0.000000	0.019057	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
14	0.000000	0.000000	0.010510	0.000000	0.000000	0.000000	0.000000	0.000000	0.011084

	0	1	2	3	4	5	6	7	8
15	0.008209	0.025194	0.008855	0.000000	0.000000	0.000000	0.008873	0.018500	0.000000
16	0.000000	0.005588	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
17	0.007261	0.003184	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
18	0.000000	0.059666	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
19	0.000000	0.032091	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
20	0.000000	0.121741	0.000000	0.000000	0.000000	0.000000	0.002656	0.000000	0.000000
21	0.000000	0.007665	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
22	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
23	0.000000	0.061158	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
24	0.000000	0.155637	0.000000	0.000000	0.000000	0.006266	0.000000	0.006250	0.000000
25	0.000000	0.018841	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
26	0.000000	0.011989	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
27	0.000000	0.055140	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
28	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
29	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...
10563	0.018679	0.045045	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10564	0.000000	0.007924	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10565	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10566	0.000000	0.028743	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10567	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10568	0.000000	0.003674	0.000000	0.000000	0.004672	0.000000	0.000000	0.000000	0.000000
10569	0.181761	0.007969	0.019608	0.000000	0.000000	0.020533	0.019647	0.040963	0.020679
10570	0.000000	0.042504	0.010458	0.010910	0.021623	0.000000	0.000000	0.000000	0.000000
10571	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.007199

	0	1	2	3	4	5	6	7	8
10572	0.000000	0.053027	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10573	0.000000	0.068526	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10574	0.000000	0.006553	0.000000	0.000000	0.000000	0.033768	0.000000	0.000000	0.000000
10575	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10576	0.000000	0.019719	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10577	0.108948	0.023884	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10578	0.050682	0.035555	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10579	0.005302	0.011623	0.014298	0.000000	0.002956	0.000000	0.000000	0.000000	0.000000
10580	0.006035	0.526554	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10581	0.000000	0.018231	0.000000	0.002753	0.000000	0.000000	0.000000	0.002756	0.000000
10582	0.000000	0.064906	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10583	0.000000	0.010373	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10584	0.000000	0.044031	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10585	0.000000	0.057172	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10586	0.000000	0.000000	0.000000	0.000000	0.010919	0.000000	0.000000	0.000000	0.000000
10587	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10588	0.004989	0.024060	0.005382	0.002807	0.002782	0.002818	0.000000	0.002811	0.002838
10589	0.000000	0.000000	0.000000	0.000000	0.000000	0.012825	0.012271	0.000000	0.012916
10590	0.000000	0.133369	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10591	0.000000	0.000000	0.000000	0.010201	0.000000	0.000000	0.000000	0.000000	0.000000
10592	0.000000	0.007422	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

10593 rows × 2000 columns

```
In [107]: # Combining dataframes
```

```
df_final = pd.concat([tfidf, claimant_encoded], axis=1, join='inner')
df_final.shape
```

Out[107]: (10593, 4204)

In [108]:
```
# Combining dataframes
df_final = pd.concat([df_final, dfqtr_encoded], axis=1, join='inner')
df_final = pd.concat([df_final, dfart_encoded], axis=1, join='inner')
df_final = pd.concat([df_final, tfidf2], axis=1, join='inner')
df_final.shape
```

Out[108]: (10593, 6295)

In [109]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df_final, df['labe
l'], test_size = 0.3, random_state = 3, stratify=df['label'])
```

In [110]:
```
# Applying various Classification algorithms without doing variable red
uctions using bag of words
accuracy_scores = np.zeros(7)

# Support Vector Classifier
svm = SVC().fit(X_train, y_train)
prediction1 = svm.predict(X_test)
accuracy_scores[0] = accuracy_score(y_test, prediction1)*100
print('Support Vector Classifier accuracy: {}%'.format(accuracy_scores[
0]))


# Logistic Regression
logis = LogisticRegression().fit(X_train, y_train)
prediction2 = logis.predict(X_test)
accuracy_scores[1] = accuracy_score(y_test, prediction2)*100
print('Logistic Regression accuracy: {}%'.format(accuracy_scores[1]))

# K Nearest Neighbors
knn = KNeighborsClassifier().fit(X_train, y_train)
prediction3 = knn.predict(X_test)
```

```
accuracy_scores[2] = accuracy_score(y_test, prediction3)*100
print('K Nearest Neighbors Classifier accuracy: {}%'.format(accuracy_scores[2]))

# Gaussian Naive Bayes
#clf = GaussianNB().fit(X_train, y_train)
#prediction4 = clf.predict(X_test)
#accuracy_scores[3] = accuracy_score(y_test, prediction4)*100
#print('Gaussian Naive Bayes Classifier accuracy: {}%'.format(accuracy_scores[3]))

# Decision Tree
#decision = DecisionTreeClassifier().fit(X_train, y_train)
#prediction4 = decision.predict(X_test)
#accuracy_scores[3] = accuracy_score(y_test, prediction4)*100
#print('Decision Tree Classifier accuracy: {}%'.format(accuracy_scores[3]))


# Random Forest
random = RandomForestClassifier().fit(X_train, y_train)
prediction5 = random.predict(X_test)
accuracy_scores[4] = accuracy_score(y_test, prediction5)*100
print('Random Forest Classifier accuracy: {}%'.format(accuracy_scores[4]))

# Gradient Boosting
GB = GradientBoostingClassifier().fit(X_train, y_train)
prediction6 = GB.predict(X_test)
accuracy_scores[5] = accuracy_score(y_test, prediction6)*100
print('Gradient Boosting Classifier accuracy: {}%'.format(accuracy_scores[5]))

'''#XGBoosting
xgb_model = xgb.XGBClassifier()
xgb_model.fit(X_train, y_train)
prediction7 = xgb_model.predict(X_test)
accuracy_scores[6] = accuracy_score(y_test, prediction7)*100
```

```
print('XGBoost Classifier accuracy: {}%'.format(accuracy_scores[6]))
'''
```

```
Support Vector Classifier accuracy: 48.7413467589679%
Logistic Regression accuracy: 63.40465701699182%
K Nearest Neighbors Classifier accuracy: 57.457520453115166%
Random Forest Classifier accuracy: 56.733794839521714%
Gradient Boosting Classifier accuracy: 63.436123348017624%
```

Out[110]:

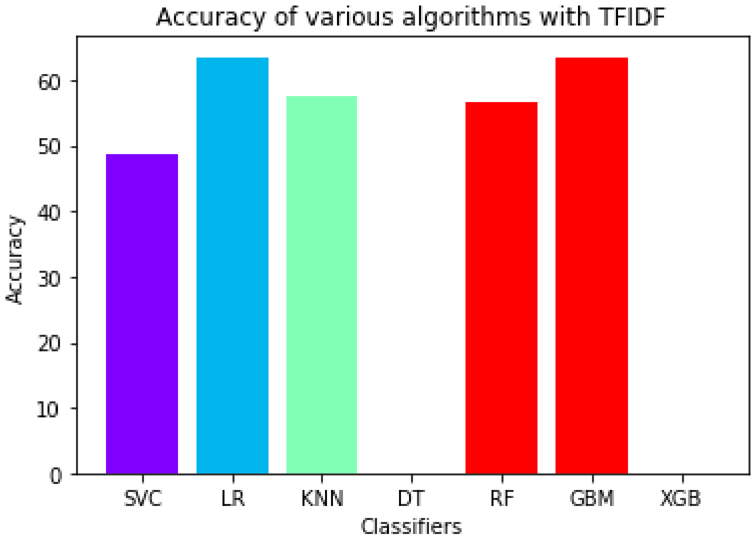
```
"#XGBoosting\nxgb_model = xgb.XGBClassifier() \nxgb_model.fit(X_train,
y_train)\nprediction7 = xgb_model.predict(X_test)\naccuracy_scores[6] =
accuracy_score(y_test, prediction7)*100\nprint('XGBoost Classifier accu
racy: {}%'.format(accuracy_scores[6]))\n"
```

In [111]:

```
# Accuracy comparison of various algorithms for Tfidf
colors = cm.rainbow(np.linspace(0, 2, 9))
labels = ['SVC', 'LR', 'KNN', 'DT', 'RF', 'GBM', 'XGB']
plt.bar(labels,
        accuracy_scores,
        color = colors)
plt.xlabel('Classifiers')
plt.ylabel('Accuracy')
plt.title('Accuracy of various algorithms with TFIDF')
```

Out[111]:

```
Text(0.5, 1.0, 'Accuracy of various algorithms with TFIDF')
```



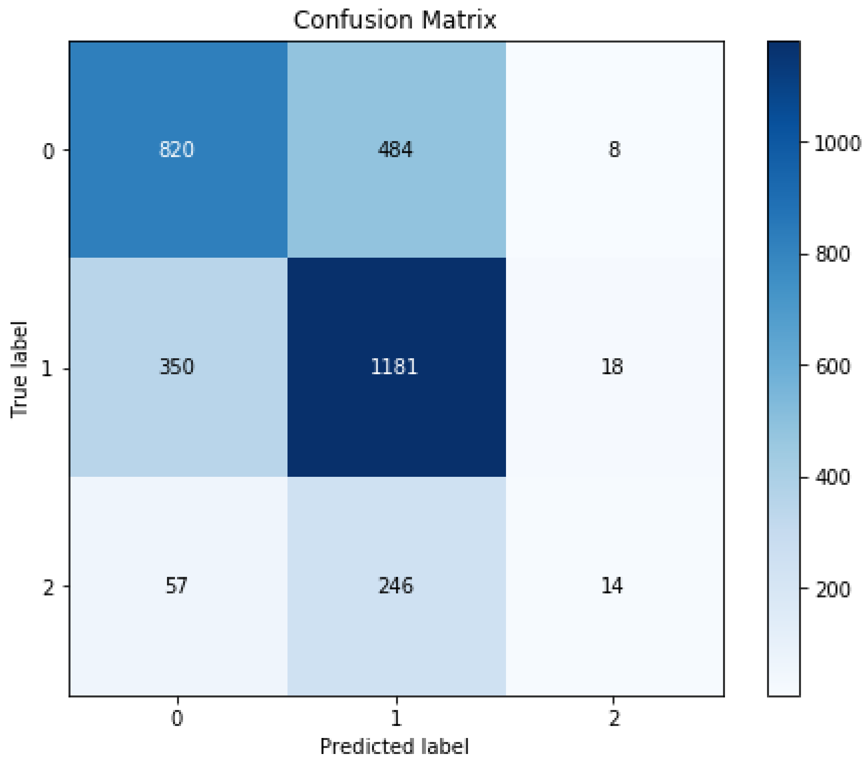
In [112]:

```
# check validation statistics (Classification Summary)
print(classification_report(y_test, prediction2)) # from confusion matrix Logistics Regression perform well
# Plot confusion Matrix
skplt.metrics.plot_confusion_matrix(y_test, prediction2, figsize=(8, 6))
plt.show()
```

```
              precision    recall  f1-score   support

           0       0.67      0.62      0.65      1312
           1       0.62      0.76      0.68      1549
           2       0.35      0.04      0.08       317

    accuracy                           0.63      3178
   macro avg       0.55      0.48      0.47      3178
weighted avg       0.61      0.63      0.61      3178
```



Based on the above, it is evident that "Partly True" and "True" news are often mis classified between themselves. Combining these two, would definitely massively improve the accuracy of the model. Another alternate would be filtering out all "opinions" and only modelling facts, which are more objective and less subjective to the decision of the person who classified the dataset initially.

Hyperparameter Tuning

In [113]:
```
# taking logistic regression as the final model (stable and higher accu
racy)
dual=[True,False]
max_iter=[100,110,120,130,140]
```

```
C = [0.001, 0.01, 0.1, 1, 10, 100, 1000]
param_grid = dict(dual=dual,max_iter=max_iter,C=C)
```

In [114]:

```
#Logsitic Regression hyperparameter tuning
from sklearn.model_selection import GridSearchCV
lr = LogisticRegression(penalty='l2')
grid = GridSearchCV(estimator=lr, param_grid=param_grid, cv = 3, n_jobs
=-1)
#Model after tuning
grid_result = grid.fit(X_train, y_train)
# Summarize results
print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_
params_))
```

```
Best: 0.645718 using {'C': 0.1, 'dual': True, 'max_iter': 100}
```

In [115]:

```
# Logistic Regression accuracy after hypertuning
prediction_logit_new = grid_result.predict(X_test)
accuracy_scores_logit_new = accuracy_score(y_test, prediction_logit_new
)*100
print('Logistic Regression accuracy after hyperparameter tuning: {}%'.f
ormat(accuracy_scores_logit_new))
```

```
Logistic Regression accuracy after hyperparameter tuning: 64.1598489616
1108%
```

SMOTE Example: Dealing with imbalanced dataset

In [118]:

```
!pip install imblearn
```

```
Collecting imblearn
  Downloading https://files.pythonhosted.org/packages/81/a7/4179e6ebfd6
54bd0eac0b9c06125b8b4c96a9d0a8ff9e9507eb2a26d2d7e/imblearn-0.0-py2.py3-
none-any.whl
Collecting imbalanced-learn (from imblearn)
  Downloading https://files.pythonhosted.org/packages/e6/62/08c14224a7e
242df2cef7b312d2ef821c3931ec9b015ff93bb52ec8a10a3/imbalanced_learn-0.5.
```

```
0-py3-none-any.whl (173kB)
Requirement already satisfied: numpy>=1.11 in c:\programdata\anaconda3
\lib\site-packages (from imbalanced-learn->imblearn) (1.16.4)
Requirement already satisfied: scipy>=0.17 in c:\programdata\anaconda3
\lib\site-packages (from imbalanced-learn->imblearn) (1.2.1)
Requirement already satisfied: scikit-learn>=0.21 in c:\programdata\ana
conda3\lib\site-packages (from imbalanced-learn->imblearn) (0.21.2)
Requirement already satisfied: joblib>=0.11 in c:\programdata\anaconda3
\lib\site-packages (from imbalanced-learn->imblearn) (0.13.2)
Installing collected packages: imbalanced-learn, imblearn
Successfully installed imbalanced-learn-0.5.0 imblearn-0.0
```

In [126]:

```
X_train.shape, y_train.shape
from imblearn.over_sampling import SMOTE
smote = SMOTE('minority')
X_sm, y_sm = smote.fit_sample(X_train, y_train)
print(X_sm.shape, y_sm.shape)
```

```
(10292, 6295) (10292,)
```

In [125]:

```
print("Number transactions X_train dataset: ", X_train.shape)
print("Number transactions y_train dataset: ", y_train.shape)
print("Number transactions X_test dataset: ", X_test.shape)
print("Number transactions y_test dataset: ", y_test.shape)
```

```
Number transactions X_train dataset:  (7415, 6295)
Number transactions y_train dataset:  (7415,)
Number transactions X_test dataset:  (3178, 6295)
Number transactions y_test dataset:  (3178,)
```

In [124]:

```
sm = SMOTE(random_state = 2)
X_train_res, y_train_res = sm.fit_sample(X_train, y_train.ravel())

print('After OverSampling, the shape of train_X: {}'.format(X_train_res
.shape))
print('After OverSampling, the shape of train_y: {} \n'.format(y_train_
res.shape))
```

```
After OverSampling, the shape of train_X: (10845, 6295)
After OverSampling, the shape of train_y: (10845,)
After OverSampling, the shape of train_y: (10845,)

After OverSampling, counts of label '1': 3615
After OverSampling, counts of label '0': 3615
```

In [129]:

```
print("After OverSampling, counts of label '1': {}".format(sum(y_train_res == 1)))
print("After OverSampling, counts of label '0': {}".format(sum(y_train_res == 0)))
print("After OverSampling, counts of label '2': {}".format(sum(y_train_res == 2)))
```

```
After OverSampling, counts of label '1': 3615
After OverSampling, counts of label '0': 3615
After OverSampling, counts of label '2': 3615
```

Due to the interest of time, this new X train is not being implemented to check accuracy score. However, it is expected that accuracy (especially that of label =2) will be improved following SMOTE. As we can see above, the training data set is now balanced.