Sentiment Analysis Twitter

Kaggle

JULY 2

IIT KANPUR Summer Internship Project
Authored by: Aman Srivastava
aman13799@gmail.com

kaggle

Tweet Polarity Classification:

Given a message, classify whether the message is of positive, negative, or neutral sentiment.

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment analysis allows businesses to identify customer sentiment toward products, brands or services in online conversations and feedback.

> ***"The meaning of your communication is the response that you get."***

INTRODUCTION

Types of Sentiment Analysis

Sentiment analysis models focus on polarity (positive, negative, neutral) but also on feelings and emotions (angry, happy, sad, etc), and even on intentions (e.g. interested v. not interested).

Here are some of the most popular types of sentiment analysis:

Fine-grained Sentiment Analysis

If polarity precision is important to your business, you might consider expanding your polarity categories to include:

Very positive
Positive

Neutral
Negative
Very negative
This is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:
Very Positive = 5 stars
Very Negative = 1 star

Emotion detection

This type of sentiment analysis aims at detecting emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithms.

One of the downsides of using lexicons is that people express emotions in different ways. Some words that typically express anger, like bad or kill (e.g. your product is so bad or your customer support is killing me) might also express happiness (e.g. this is bad ass or you are killing it).

Aspect-based Sentiment Analysis

Usually, when analyzing sentiments of texts, let's say product reviews, you'll want to know which aspects or features people are mentioning in a positive, neutral, or negative way. That's where aspect-based sentiment analysis can help, for example in this text: "The battery life of this camera is too short", an aspect-based classifier would be able to determine that the sentence expresses a negative opinion about the feature battery life.

Multilingual sentiment analysis

Multilingual sentiment analysis can be difficult. It involves a lot of preprocessing and resources. Most of these resources are available online (e.g. sentiment lexicons), while others need to be created (e.g. translated corpora or noise detection algorithms), but you'll need to know how to code to use them.

OUR APPROACH

After exploratory analysis of the given dataset it was found that the dataset was imbalanced. So, after cleaning the data, I applied data augmentation to make dataset balanced and then trained a Neural Network on it.

Cleaning the dataset:

Tweets are noisy data and the sentences may contain many different types of short hand words instead of complete words. This is maily because tweets have a maximum limit of 250 charecters. So people try to use shorthand words/emoticons as much as possible to fit the tweets under the max limit.

Also, tweets contain hyperlinks which should be removed as they are of little use to us.

- Importing the dataset from kaggle

```
In [2]: import os
        print(os.listdir("input"))

        ['test.tsv', 'train.tsv']

In [3]: # Loading Data
        train_df = pd.read_csv("input/train.tsv",sep='\t')
        test_df = pd.read_csv("input/test.tsv",sep='\t')

In [4]: #Training Data Set
        train_df.head(10)
```

Out[4]:

	tweet_id	sentiment	tweet_text
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...

In [5]:
```
#Testing Data Set
test_df.head()
print('Testing data set has no Label column')
print(test_df.head(10))
```

```
Testing data set has no Label column
             tweet_id                                         tweet_text
0  264238274963451904  @jjuueellzz down in the Atlantic city, ventnor...
1  218775148495515649  Musical awareness: Great Big Beautiful Tomorro...
2  258965201766998017  On Radio786 100.4fm 7:10 Fri Oct 19 Labour ana...
3  262926411352903682  Kapan sih lo ngebuktiin,jan ngomong doang Susa...
4  171874368908050432  Excuse the connectivity of this live stream, f...
5  256010056042903296  Show your LOVE for your local field & it might...
6  253809989599232000  Milton on Bolton Wanderers 2 v 2 Leeds United,...
7  261776619146985472  @firecore Can you tell me when an update for t...
8  264143999374356481  @Heavensbasement The Crown, Filthy McNastys, K...
9  223052929131757571  Uncover the Eternal City! Return flights to Ro...
```

In [6]:
```
# Training Data Set Information
print("Training Data Set Info - Total Rows | Total Columns | Total Null Values")
print(train_df.info())
```

```
Training Data Set Info - Total Rows | Total Columns | Total Null Values
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21630 entries, 0 to 21629
Data columns (total 3 columns):
tweet_id      21630 non-null int64
sentiment     21630 non-null object
tweet_text    21630 non-null object
dtypes: int64(1), object(2)
memory usage: 507.1+ KB
None
```

- We now merge both train and test dataset as both dataset needs to be cleaned.

In [8]:
```
# Merging both the data sets as tweets in both the data set is unstructured
combine_df = train_df.append(test_df, ignore_index = True, sort = False)
combine_df.head()
```

Out[8]:

	tweet_id	sentiment	tweet_text
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...

In [9]:
```
# Combine (Merged) Data Set Information
print("Combine Data Set Info - Total Rows | Total Columns | Total Null Values")
print(combine_df.info())
```

```
Combine Data Set Info - Total Rows | Total Columns | Total Null Values
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27028 entries, 0 to 27027
Data columns (total 3 columns):
tweet_id      27028 non-null int64
sentiment     21630 non-null object
tweet_text    27028 non-null object
dtypes: int64(1), object(2)
memory usage: 633.6+ KB
None
```

In [10]:
```
# Importing HTMLParser
from html.parser import HTMLParser
html_parser = HTMLParser()
```

In [11]:
```
# Created a new columns i.e. clean_tweet contains the same tweets but cleaned version
combine_df['clean_tweet'] = combine_df['tweet_text'].apply(lambda x: html_parser.unescape(x))
combine_df.head(10)
```

- **Next we wrote a simple function which removes pattern from the dataset by passing regex pattern and text into it.**

```
In [13]: def remove_pattern(input_txt, pattern):
             r = re.findall(pattern, input_txt)
             for i in r:
                 input_txt = re.sub(i, '', input_txt)
             return input_txt
```

```
In [21]: # remove twitter handles (@user)
         combine_df['clean_tweet'] = np.vectorize(remove_pattern)(combine_df['clean_tweet'], "@[\w]*")
         combine_df.head(100)
```

Out[21]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	Gas by my house hit $3.39!!!! I'm going to Cha...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	Theo Walcott is still shit, watch Rafa and Joh...
2	262163168678248449	negative	Its not that I'm a GSP fan, I just hate Nick D...	Its not that I'm a GSP fan, I just hate Nick D...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	Iranian general says Israel's Iron Dome can't ...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	Tehran, Mon Amour: Obama Tried to Establish Ti...
...	...	...	...	...
95	263132960507703296	neutral	and the 4th one is for Harry Styles!! <33333	and the 4th one is for Harry Styles!! <33333
96	264229531773177856	neutral	this guy so much like bruno mars. it's crazy. ...	this guy so much like bruno mars. it's crazy. ...
97	259022549604790273	neutral	Blog Post: MTV's Teen Mom 2 Returns for an I...	Blog Post: MTV's Teen Mom 2 Returns for an I...
98	253050608322502657	neutral	The Business Leader's Award ceremony will be h...	The Business Leader's Award ceremony will be h...
99	262999273275260928	positive	@elifyk1 Hi Elf Greetings from USA midwest ...	Hi Elf Greetings from USA midwest How is s...

100 rows × 4 columns

```
In [22]: combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: x.lower())
         combine_df.head(10)
```

Out[22]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit $3.39!!!! i'm going to cha...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit, watch rafa and joh...
2	262163168678248449	negative	Its not that I'm a GSP fan, I just hate Nick D...	its not that i'm a gsp fan, i just hate nick d...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel's iron dome can't ...

We apply the following modifications to tweet:

- **Remove Usernames (@user)**
- **Convert tweets to lowercase**
- **Converting apostrophe words to set of words. (“I’m” to “I” am etc.) using a apostrophe words dictionary**
- **Remove hyperlinks**
- **Convert short words to complete words using a short word dictionary**
- **Converting emoticons to word emotions using a emoticon dictionary**
- **Removing all special characters**
- **Removing all numbers**
- **Removing single characters (word length < 1)**

In [24]:
```
def lookup_dict(text, dictionary):
    for word in text.split():
        if word.lower() in dictionary:
            if word.lower() in text.split():
                text = text.replace(word, dictionary[word.lower()])
    return text
```

In [25]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: lookup_dict(x,apostrophe_dict))
combine_df.head(10)
```

Out[25]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit $3.39!!!! i am going to ch...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit, watch rafa and joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...	its not that i am a gsp fan, i just hate nick ...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel's iron dome cannot...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	tehran, mon amour: obama tried to establish ti...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...	i sat through this whole movie just for harry ...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...	with j davlar 11th. main rivals are team polan...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...	talking about act's && sat's, deciding where i...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...	why is happy valentines day trending? it has...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...	they may have a superbowl in dallas, but dalla...

In [26]:
```
def remove_site(text):
    URLless_string = re.sub(r'\w+:\/{2}[\d\w-]+(\.[\d\w-]+)*(?:(?:\/[^\s/]*))*', '', text)
    return URLless_string
```

In [29]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: remove_site(x))
combine_df.head(10)
```

In [31]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: lookup_dict(x,short_word_dict))
combine_df.head(10)
```

Out[31]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit $3.39!!!! i am going to ch...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit, watch rafa and joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...	its not that i am a gsp fan, i just hate nick ...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel's iron dome cannot...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	tehran, mon amour: obama tried to establish ti...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...	i sat through this whole movie just for harry ...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...	with j davlar 11th. main rivals are team polan...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...	talking about act's && sat's, deciding where i...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...	why is happy valentines day trending? it has...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...	they may have a superbowl in dallas, but dalla...

In [33]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: lookup_dict(x,emoticon_dict))
combine_df.head(10)
```

Out[33]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit $3.39!!!! I am going to ch...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit, watch rafa and joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...	its not that I am a gsp fan, i just hate nick ...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel's iron dome cannot...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	tehran, mon amour: obama tried to establish ti...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...	i sat through this whole movie just for harry ...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...	with j davlar 11th. main rivals are team polan...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...	talking about act's && sat's, deciding where i...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...	why is happy valentines day trending? it has...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...	they may have a superbowl in dallas, but dalla...

In [34]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: re.sub(r'[^\w\s]',' ',x))
combine_df.head(10)
```

Out[34]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit 3 39 I am going to ch...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit watch rafa and joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...	its not that I am a gsp fan i just hate nick ...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel s iron dome cannot...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	tehran mon amour obama tried to establish ti...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...	i sat through this whole movie just for harry ...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...	with j davlar 11th main rivals are team polan...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...	talking about act s sat s deciding where i...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...	why is happy valentines day trending it has...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...	they may have a superbowl in dallas but dalla...

In [35]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: re.sub(r'[^a-zA-Z0-9]',' ',x))
combine_df.head(10)
```

Out[35]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit 3 39 I am going to ch...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit watch rafa and joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...	its not that I am a gsp fan i just hate nick ...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel s iron dome cannot...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	tehran mon amour obama tried to establish ti...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...	i sat through this whole movie just for harry ...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...	with j davlar 11th main rivals are team polan...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...	talking about act s sat s deciding where i...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...	why is happy valentines day trending it has...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...	they may have a superbowl in dallas but dalla...

In [36]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: re.sub(r'[^a-zA-Z]',' ',x))
combine_df.head(10)
```

Out[36]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit I am going to ch...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit watch rafa and joh...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...	its not that I am a gsp fan i just hate nick ...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel s iron dome cannot...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	tehran mon amour obama tried to establish ti...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...	i sat through this whole movie just for harry ...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...	with j davlar th main rivals are team polan...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...	talking about act s sat s deciding where i...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...	why is happy valentines day trending it has...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...	they may have a superbowl in dallas but dalla...

In [37]:
```
combine_df['clean_tweet'] = combine_df['clean_tweet'].apply(lambda x: ' '.join([w for w in x.split() if len(w)>1]))
combine_df.head(10)
```

Out[37]:

	tweet_id	sentiment	tweet_text	clean_tweet
0	264183816548130816	positive	Gas by my house hit $3.39!!!! I'm going to Cha...	gas by my house hit am going to chapel hill on...
1	263405084770172928	negative	Theo Walcott is still shit, watch Rafa and Joh...	theo walcott is still shit watch rafa and john...
2	262163168678248449	negative	its not that I'm a GSP fan, i just hate Nick D...	its not that am gsp fan just hate nick diaz ca...
3	264249301910310912	negative	Iranian general says Israel's Iron Dome can't ...	iranian general says israel iron dome cannot d...
4	262682041215234048	neutral	Tehran, Mon Amour: Obama Tried to Establish Ti...	tehran mon amour obama tried to establish ties...
5	264229576773861376	neutral	I sat through this whole movie just for Harry ...	sat through this whole movie just for harry an...
6	264105751826538497	positive	with J Davlar 11th. Main rivals are team Polan...	with davlar th main rivals are team poland hop...
7	264094586689953794	negative	Talking about ACT's && SAT's, deciding where I...	talking about act sat deciding where want to g...
8	212392538055778304	neutral	Why is Happy Valentines Day trending? It's o...	why is happy valentines day trending it has it...
9	254941790757601280	negative	They may have a SuperBowl in Dallas, but Dalla...	they may have superbowl in dallas but dallas a...

We then save the processed dataset as a csv.

```
In [11]: traindata=pd.read_csv('project/input/train_preprocessed.csv')
```

```
In [96]: t_id = pd.read_csv("project/input/test_samples.csv")
         test = pd.read_csv('project/input/train_preprocessed.csv')
         test = test[['clean_tweet']]
         test = test[21630:]
         test = test[['clean_tweet']]
         test = test.reset_index(drop=True)
         t_id = t_id[['tweet_id']]
         t_id = t_id.reset_index(drop=True)
         testdata = pd.concat([t_id, test], axis = 1)
```

```
In [12]: traindata.head()
```

Out[12]:

	Unnamed: 0	clean_tweet	sentiment
0	0	gas by my house hit 4m going to chapel hill on...	positive
1	1	theo walcott is still shit watch rafa and john...	negative
2	2	its not that im gsp fan just hate nick diaz ca...	negative
3	3	iranian general says israel iron dome cannot d...	negative
4	4	tehran mon amour obama tried to establish ties...	neutral

```
In [49]: testdata.head()
```

Out[49]:

	tweet_id	clean_tweet
0	264238274963451904	down in the atlantic city ventnor margate ocea...
1	218775148495515649	musical awareness great big beautiful tomorrow...
2	258965201766998017	on radio fm fri oct labour analyst shawn hatti...
3	262926411352903682	kapan sih lo ngebuktiin jan ngomong doang susa...
4	171874368908050432	excuse the connectivity of this live stream fr...

```
In [23]: traindata=traindata[['clean_tweet','sentiment']]
         traindata = traindata[:21629]
```

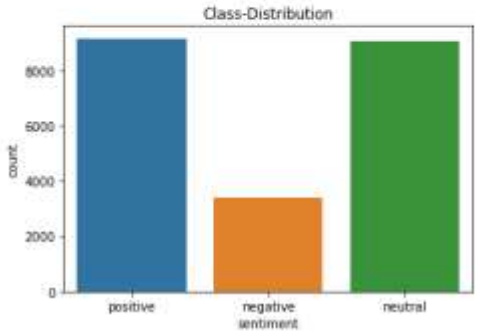
As we can see that the value counts of negative sentiments is a lot less than positive and negative sentiments. This makes the dataset heavily imbalanced.

```
In [14]: traindata['sentiment'].value_counts()
```

```
Out[14]: positive    9155
         neutral     9075
         negative    3400
         Name: sentiment, dtype: int64
```

```
In [15]: %matplotlib inline
         sns.countplot(traindata['sentiment'])
         plt.title('Class-Distribution')
```

```
Out[15]: Text(0.5, 1.0, 'Class-Distribution')
```



Data Augmentation using nlpaug

In [16]:
```
import nlpaug.augmenter.sentence as nas
```

In [17]:
```
WANDB_API_KEY='aman'
```

In [18]:
```
aug = nas.ContextualWordEmbsForSentenceAug(model_path='xlnet-base-cased')
```

```
HBox(children=(IntProgress(value=0, description='Downloading', max=798011, style=ProgressStyle(description_wid…


HBox(children=(IntProgress(value=0, description='Downloading', max=760, style=ProgressStyle(description_width=…


HBox(children=(IntProgress(value=0, description='Downloading', max=467042463, style=ProgressStyle(description_…
```

In [26]:
```
ls=[]
def data_augument(df):
    augmented_texts = aug.augment(df, n=2)
    for i in augmented_texts:
        ls.append(i)
    return(augmented_texts)
```

In [27]:
```
from tqdm._tqdm_notebook import tqdm_notebook
tqdm_notebook.pandas()
```

```
/home/shriaman/anaconda3/lib/python3.7/site-packages/tqdm/std.py:648: FutureWarning: The Panel class is removed from pandas. Accessing it from the top-level namespace will also be removed in the next version
  from pandas import Panel
```

In [28]:
```
traindata[traindata['sentiment']=='negative']['clean_tweet'].progress_apply(data_augument)
```

```
HBox(children=(IntProgress(value=0, max=3400), HTML(value='')))
```

Out[28]:
```
1        [theo walcott is still shit watch rafa and joh...
2        [its not that am gsp fan just hate nick diaz c...
3        [iranian general says israel iron dome cannot ...
7        [talking about act sat deciding where want to ...
9        [they may have superbowl in dallas but dallas ...
                               ...
21588    [sept reuters jpmorgan chase co said it had no...
21593    [that th down play in the colts game back in i...
21596    [ran into alonso at the gym literally have not...
21610    [green bay drafted rogers at about the same sp...
21611    [aaah st fixture list saw earlier had valencia...
Name: clean_tweet, Length: 3400, dtype: object
```

In [29]:
```
array=np.array(ls)
```

In [30]:
```
np.save('array',array)
```

In [31]:
```
augmented_texts=np.load('array.npy')
```

In [32]:
```
aug_data=pd.DataFrame(augmented_texts,columns=['clean_tweet'])
```

In [33]:
```
aug_data['sentiment']='negative'
```

In [34]:
```
aug_data.head()
```

Out[34]:

	clean_tweet	sentiment
0	theo walcott is still shit watch rafa and john...	negative
1	theo walcott is still shit watch rafa and john...	negative
2	its not that am gsp fan just hate nick diaz ca...	negative
3	its not that am gsp fan just hate nick diaz ca...	negative
4	iranian general says israel iron dome cannot d...	negative

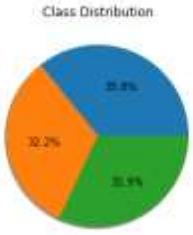
```
In [35]: traindata=pd.concat([traindata, aug_data], join="outer").sample(frac=1).reset_index(drop=True)

In [36]: traindata.head()
```

Out[36]:

	clean_tweet	sentiment
0	okay tomorrow is hell day cannot wait to get i...	negative
1	what is tree favourite day of the christian ca...	neutral
2	need some advice how can spin the wiiu loss now...	negative
3	my shuffle is favoring mariah carey hardcore t...	positive
4	lakers notebook nash listed as doubtful for sh...	negative

After augmnetation the dataset is now balanced

```
In [37]: plt.pie(traindata['sentiment'].value_counts(), autopct='%1.1f%%', shadow=True)
plt.title('Class Distribution');
plt.show()
```



After text augmentation we remove stop words and lemmatize the dataset again as augmented tweets might have stop words in them.

```
In [38]: train_process=traindata.copy()

In [42]: stop_words = stopwords.words('english')
train_process['clean_tweet'] = train_process['clean_tweet'].apply(lambda x: " ".join(x for x in x.split() if x not in stop_words

In [97]: testdata['clean_tweet'] = testdata['clean_tweet'].apply(lambda x: " ".join(x for x in x.split() if x not in stop_words))
```

Lemmatize the sentences

```
In [45]: train_process['tweet_lemma']=train_process['clean_tweet'].progress_apply(lambda words: ' '.join( [WordNetLemmatizer().lemmatize(
train_process[['tweet_lemma','clean_tweet']].head()
```

```
HBox(children=(IntProgress(value=0, max=28489), HTML(value='')))
```

Out[45]:

	tweet_lemma	clean_tweet
0	okay tomorrow hell day cannot wait get car dri...	okay tomorrow hell day cannot wait get car dri...
1	tree favourite day christian calendar ash wednesday	tree favourite day christian calendar ash wedn...
2	need advice spin wiiu loss mtsu beat bowling gr...	need advice spin wiiu loss mtsu beat bowling gr...
3	shuffle favoring mariah carey hardcore today w...	shuffle favoring mariah carey hardcore today w...
4	lakers notebook nash listed doubtful showdown ...	lakers notebook nash listed doubtful showdown ...

```
In [98]: testdata['tweet_lemma']=testdata['clean_tweet'].progress_apply(lambda words: ' '.join( [WordNetLemmatizer().lemmatize(w) for w i
testdata[['tweet_lemma','clean_tweet']].head()
```

```
HBox(children=(IntProgress(value=0, max=5398), HTML(value='')))
```

Out[98]:

	tweet_lemma	clean_tweet
0	atlantic city ventnor margate ocean city area ...	atlantic city ventnor margate ocean city area ...
1	musical awareness great big beautiful tomorrow...	musical awareness great big beautiful tomorrow...
2	radio fm fri oct labour analyst shawn hattingh...	radio fm fri oct labour analyst shawn hattingh...
3	kapan sih lo ngebuktiin jan ngomong doang susa...	kapan sih lo ngebuktiin jan ngomong doang susa...
4	excuse connectivity live stream baba amr many ...	excuse connectivity live stream baba amr many ...

```
In [61]: train_process['sentiment']
```

Out[61]:

```
0       negative
1        neutral
2       negative
3       positive
4       negative
```

Next, we convert sentiment values into numbers.

In [62]:
```
def mark_sentiment(sentiment):
    if(sentiment=='positive'):
        return 4
    if(sentiment=='negative'):
        return 0
    else:
        return 2
train_process['sentiment']=train_process['sentiment'].progress_apply(mark_sentiment)
train_process['sentiment']
```

```
HBox(children=(IntProgress(value=0, max=28409), HTML(value='')))
```

Out[62]:
```
0        0
1        2
2        0
3        4
4        0
        ..
28404    4
28405    0
28406    4
28407    0
28408    4
Name: sentiment, Length: 28409, dtype: int64
```

In [85]:
```
# convert text into numeric values using tf-idf vectorizer
tfidfconverter = TfidfVectorizer(
                                 min_df=5,
                                 max_df=0.7,
                                 ngram_range=(1,2),
                                 stop_words='english'
                                )
X = tfidfconverter.fit_transform(train_process['tweet_lemma']).toarray()
y = train_process['sentiment'].values
```

Deep learning Model

In [111]:
```
ycat=pd.get_dummies(train_process['sentiment']).values
X=train_process['tweet_lemma'].values
tk = Tokenizer()
tk.fit_on_texts(X)
X_seq = tk.texts_to_sequences(X)
X_pad = pad_sequences(X_seq, maxlen=25, padding='post')
X_train, X_test, y_train, y_test = train_test_split(X_pad, ycat, test_size = 0.25, random_state = 1)
```

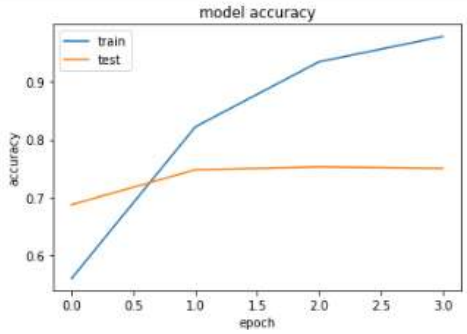
In [114]:
```
vocabulary_size = len(tk.word_counts.keys())+1
max_words = 25
embedding_size = 32
model = Sequential()
model.add(Embedding(vocabulary_size, embedding_size, input_length=max_words))
model.add(Flatten())
model.add(Dense(3, activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
model.summary()
```

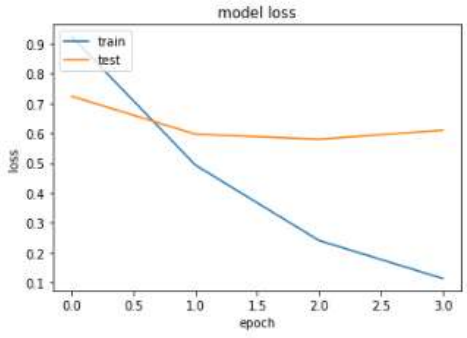
```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 25, 32)            1045856
_________________________________________________________________
flatten_2 (Flatten)          (None, 800)               0
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 2403
=================================================================
Total params: 1,048,259
Trainable params: 1,048,259
Non-trainable params: 0
_________________________________________________________________
```

In [115]:
```
history=model.fit(X_train,y_train,validation_data=(X_test,y_test),batch_size=32,epochs=4,verbose=True)
```

```
WARNING:tensorflow:From /home/shriaman/anaconda3/lib/python3.7/site-packages/keras/backend/tensorflow_backend.py:422: The name
tf.global_variables is deprecated. Please use tf.compat.v1.global_variables instead.

Train on 21306 samples, validate on 7103 samples
Epoch 1/4
21306/21306 [==============================] - 5s 255us/step - loss: 0.9237 - accuracy: 0.5606 - val_loss: 0.7237 - val_accurac
y: 0.6877
Epoch 2/4
21306/21306 [==============================] - 5s 257us/step - loss: 0.4932 - accuracy: 0.8221 - val_loss: 0.5968 - val_accurac
y: 0.7480
Epoch 3/4
21306/21306 [==============================] - 6s 261us/step - loss: 0.2403 - accuracy: 0.9344 - val_loss: 0.5798 - val_accurac
y: 0.7529
Epoch 4/4
21306/21306 [==============================] - 5s 255us/step - loss: 0.1129 - accuracy: 0.9781 - val_loss: 0.6100 - val_accurac
y: 0.7504
```

In [116]:
```
# summarize history for accuracy
plt.plot(history.history['accuracy'])
plt.plot(history.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
# summarize history for loss
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
```





The trained model obtained an accuracy of 75% on verification set. We now run the model on test dataset to upload the output of model on Kaggle.

```
In [100]: test = tfidfconverter.transform(testdata['tweet_lemma']).toarray()

In [102]: test_predictions = text_classifier.predict(test)

[Parallel(n_jobs=12)]: Using backend ThreadingBackend with 12 concurrent workers.
[Parallel(n_jobs=12)]: Done  26 tasks      | elapsed:    0.1s
[Parallel(n_jobs=12)]: Done 176 tasks      | elapsed:    0.4s
[Parallel(n_jobs=12)]: Done 400 out of 400 | elapsed:    1.1s finished

In [103]: test_predictions

Out[103]: array([4, 4, 2, ..., 2, 0, 4])

In [105]: def class_sentiment(prediction):
              n = len(prediction)
              sentiment = []
              for i in range(n):
                  if(prediction[i] == 4):
                      sentiment.append('positive')
                  elif(prediction[i]==2):
                      sentiment.append('neutral')
                  else:
                      sentiment.append('negative')
              return sentiment
          sentiment = class_sentiment(test_predictions)
          testdata['sentiment'] = sentiment
          testdata['sentiment'].value_counts()

Out[105]: neutral     3223
          positive    1715
          negative     460
          Name: sentiment, dtype: int64

In [108]: testdata = testdata[['tweet_id','sentiment']]
          test_list = []
          heading = ['tweet_id', 'sentiment']
          test_list.append(heading)
          for i in range(len(testdata['tweet_id'])):
              sub = []
              sub.append(testdata['tweet_id'][i])
              sub.append(testdata['sentiment'][i])
              test_list.append(sub)

In [110]: import csv
          with open('junelast.csv', 'w', newline='') as fp:
              a = csv.writer(fp, delimiter = ",")
              data = test_list
              a.writerows(data)
          check = pd.read_csv("junelast.csv")
          check.info()

          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 5398 entries, 0 to 5397
          Data columns (total 2 columns):
          tweet_id     5398 non-null int64
          sentiment    5398 non-null object
          dtypes: int64(1), object(1)
          memory usage: 84.5+ KB
```

On Kaggle, this approach of test augmentation and a simple neural network obtained an F1 score of .63

We can try to improve the model further by changing the model to a RNN or try with randomForrest.