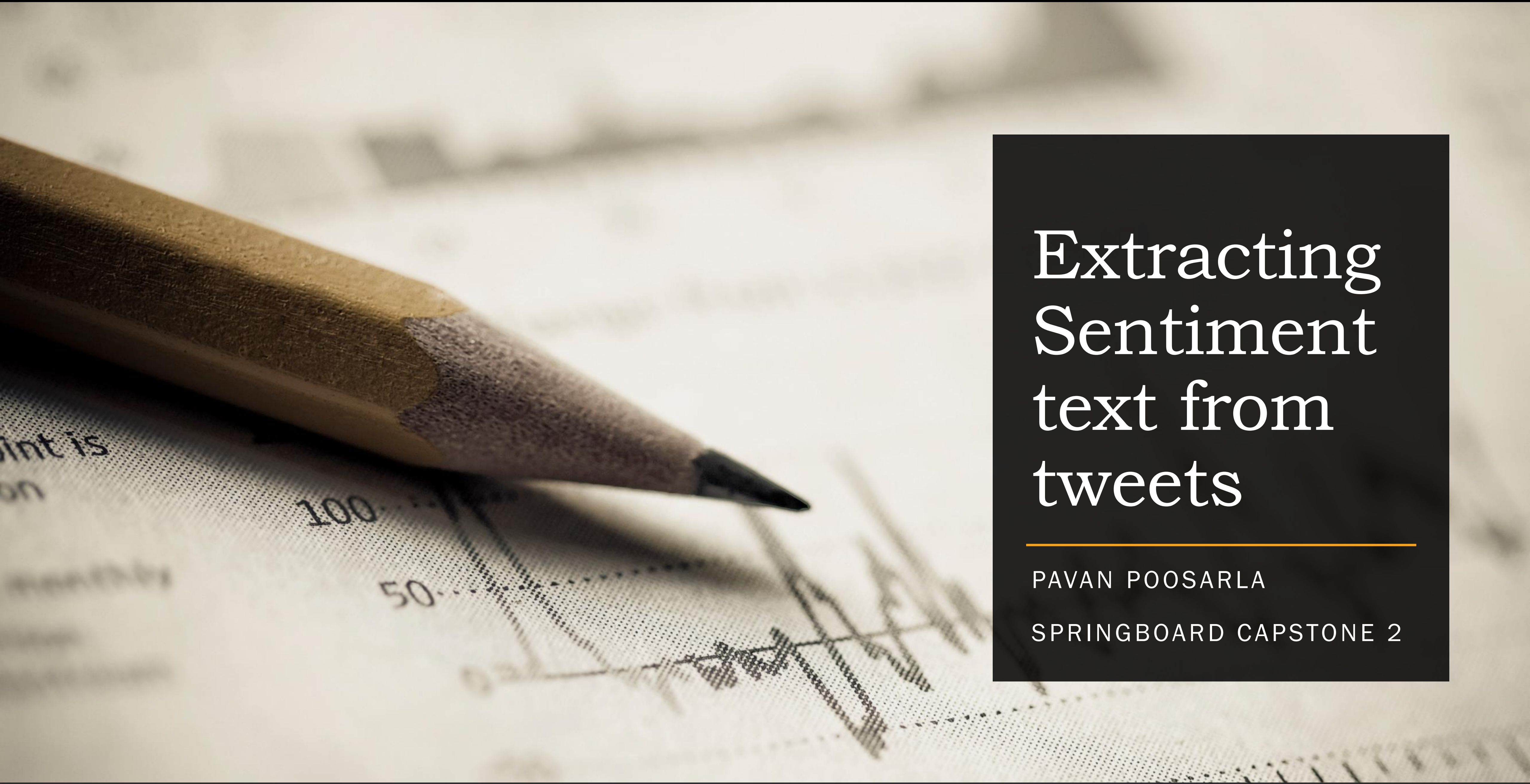


Extracting Sentiment text from tweets

PAVAN POOSARLA

SPRINGBOARD CAPSTONE 2

Problem Statement

Extract the best phrases related to the sentiment from a tweet, knowing the sentiment of the overall tweet.

Data Set : More than 25000 tweets and their sentiments (either positive, negative or neutral) along ith the selected texts from the tweets that represents the sentiment

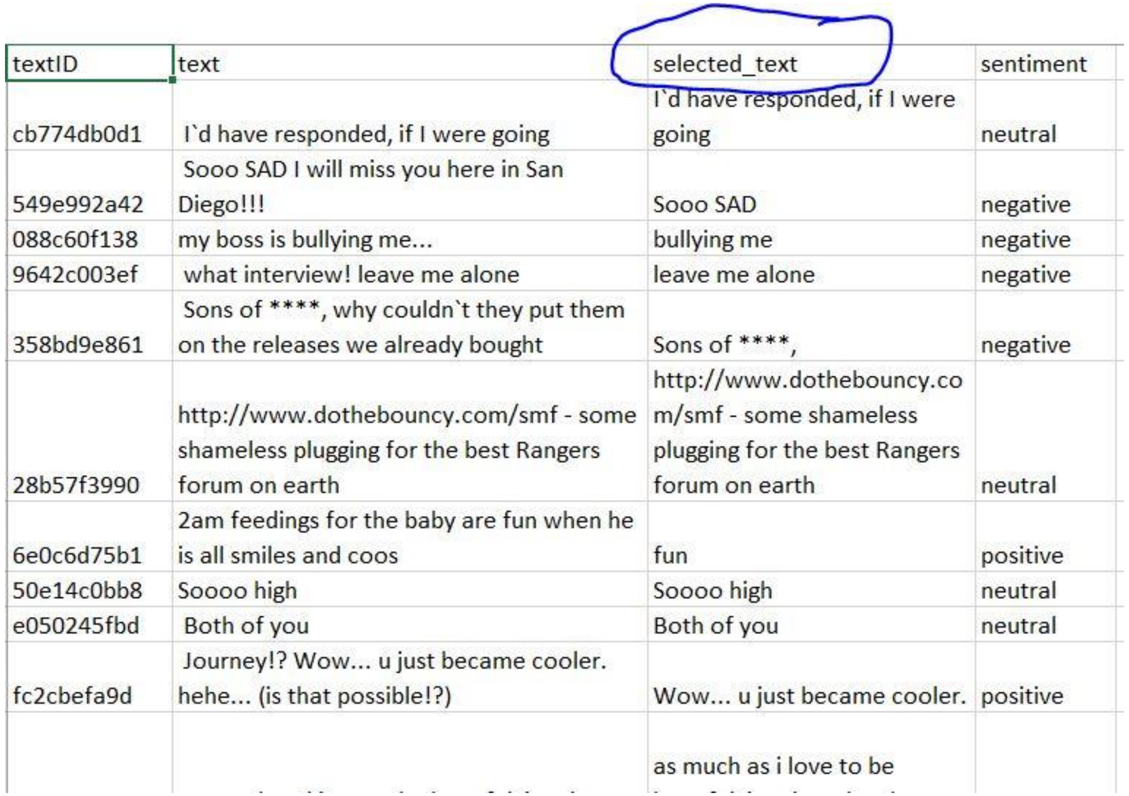
Source : https://www.kaggle.com/c/tweet-sentiment-extraction/data

Data Set

Text fields
textID : Unique ID for each tweet
Text : total text of the tweet. All in English. Upto 140 characters

Selected_text : excerpt from the tweet signifying the sentiment. This varies from a single word to the entire tweet.

Sentiment : Sentiment of the overall tweet. Can be positive, negative or neutral



textID	text	selected_text	sentiment
cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral
549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative
088c60f138	my boss is bullying me...	bullying me	negative
9642c003ef	what interview! leave me alone	leave me alone	negative
358bd9e861	Sons of ****, why couldn`t they put them on the releases we already bought	Sons of ****,	negative
28b57f3990	http://www.dothebouncy.com/smf - some shameless plugging for the best Rangers forum on earth	http://www.dothebouncy.com/smf - some shameless plugging for the best Rangers forum on earth	neutral
6e0c6d75b1	2am feedings for the baby are fun when he is all smiles and coos	fun	positive
50e14c0bb8	Soooo high	Soooo high	neutral
e050245fbd	Both of you	Both of you	neutral
fc2cbefa9d	Journey!? Wow... u just became cooler. hehe... (is that possible!?)	Wow... u just became cooler.	positive
		as much as i love to be	

Approach

Step 1 : Data Wrangling

Step 2: Exploratory Data Analysis

Step 3: Modeling and Prediction

a. Bag-of-words approach (Logistic Regression)
b. Parse tress approach (Extract relevant noun-chunks)
c. Deep Learning approach (DistilBERT)

Exploratory Data Analysis

Observations

1. Distribution of character lengths in tweets is common for all sentiments
2. Character counts in selected text tend to be much lower than full tweet for negative and positive sentiments
3. For tweets with neutral sentiment, actual tweets and selected tweets have similar length distribution

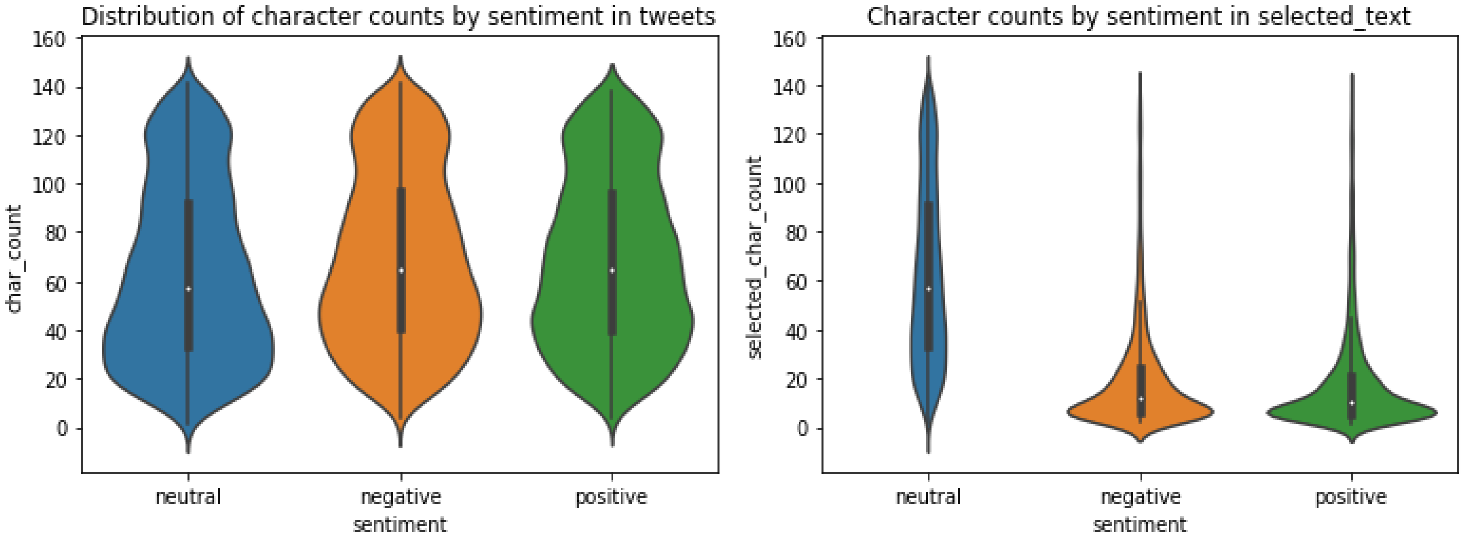
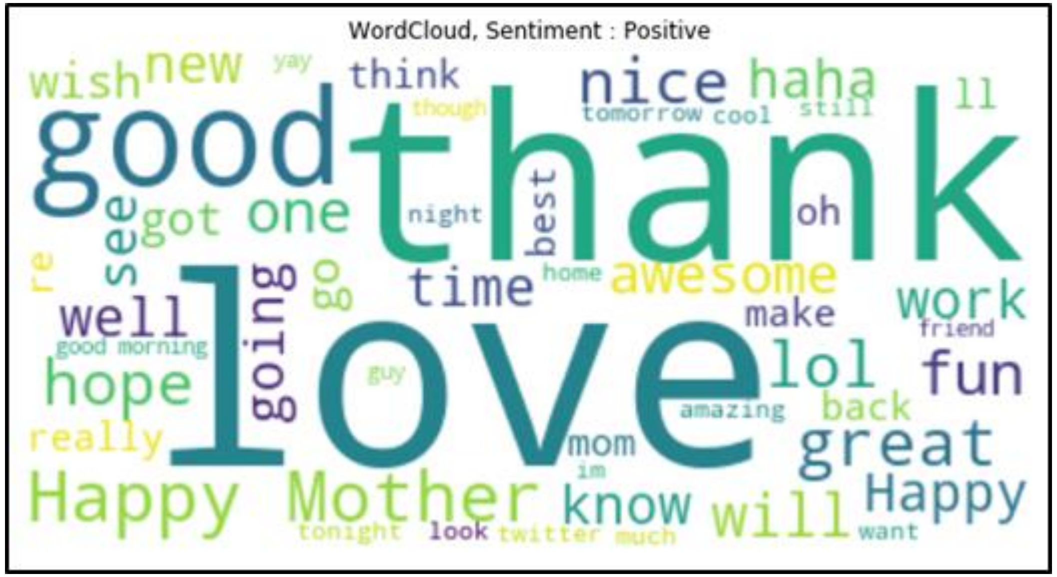
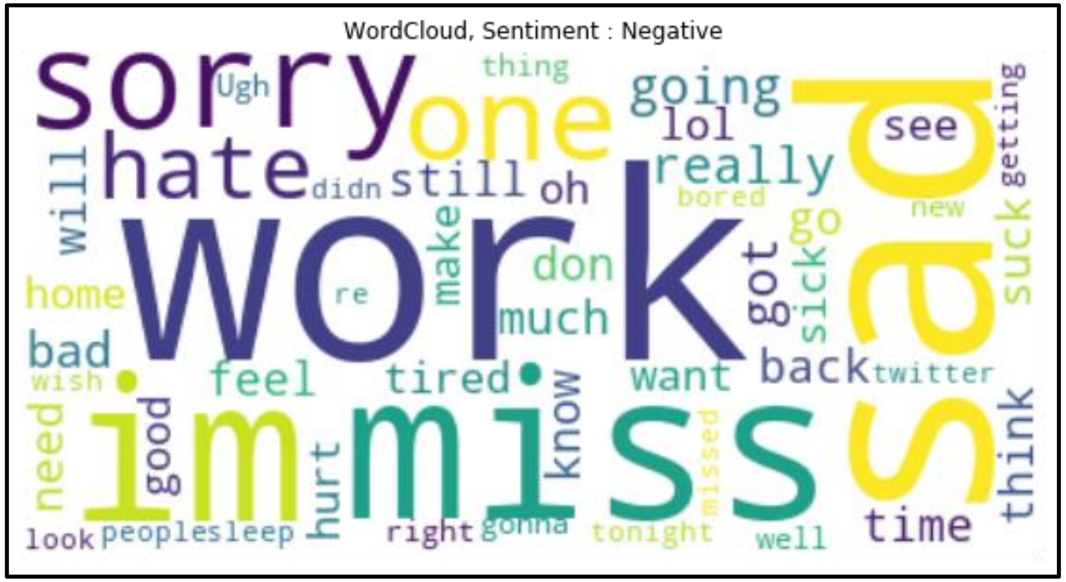


Figure Distribution of character lengths by sentiment in the full tweet and in the selected text

Text Analysis

Word clouds corresponding to different sentiments show that even individual words can show differentiation between sentiments


WordCloud, Sentiment : Positive


WordCloud, Sentiment : Negative

Modelling and Prediction

Bag-of-Words Approach

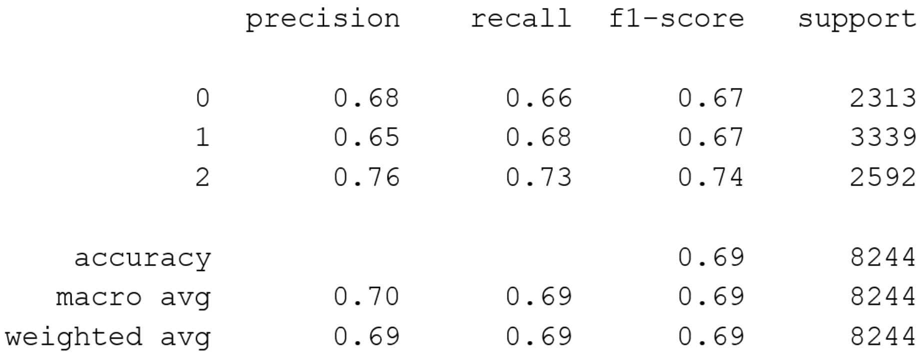
Goal : Main goal of this approach is to identify the word in the tweet with the strongest probability of the sentiment.

To get probability of sentiments for each word, we do as follows

1. Train a logistic regression sentiment classification model from the training dataset

2. Get prediction probabilities of each of sentiment for each single word

Classification Report of Logistic Regression Sentiment Classification Model



```
              precision    recall  f1-score   support

           0       0.68      0.66      0.67      2313
           1       0.65      0.68      0.67      3339
           2       0.76      0.73      0.74      2592

    accuracy                           0.69      8244
   macro avg       0.70      0.69      0.69      8244
weighted avg       0.69      0.69      0.69      8244
```

Post processing

Use model trained in previous section to get probabilities of each sentiment for each word

In each tweet, find the word with highest probability of said sentiment

Extract selected text by multiple approaches as described

The closeness of *predicted 'selected text'* and *'selected text'* is found using 'Jaccard score'

Version	Approach	**Mean Jaccard score**
Trivial score	Return entire tweet as prediction	0.589
1	Select best word in each tweet for the sentiment	0.240
2	For 'neutral' sentiment, return full tweet. Return best predictive word for rest	0.593
3	Return complete tweet for 'neutral' sentiment or if original tweet contains fewer than 6 words	0.605

Classification Report of Logistic Regression Sentiment Classification Model

```
              precision    recall  f1-score   support

           0       0.68      0.66      0.67      2313
           1       0.65      0.68      0.67      3339
           2       0.76      0.73      0.74      2592

    accuracy                           0.69      8244
   macro avg       0.70      0.69      0.69      8244
weighted avg       0.69      0.69      0.69      8244
```

Parse Trees Approach

Main drawback of bag-of-trees approach is that its performance is poor on tweets where more than one word is to be selected

We can use parse trees to select a phrase around the most predictive word and check the performance

The performance is not good because improvement gained on tweets with larger selected texts is lost for tweets with one word as selection

Version	Approach	**Mean Jaccard score**
Trivial score	Return entire tweet as prediction	0.589
4	Use parse trees to select the noun chunk containing the best predictive word and return it as predicted selection	0.584
5	Return a 5 word interval around the best predictive word as predicted selection.	0.589

Deep Learning Approach

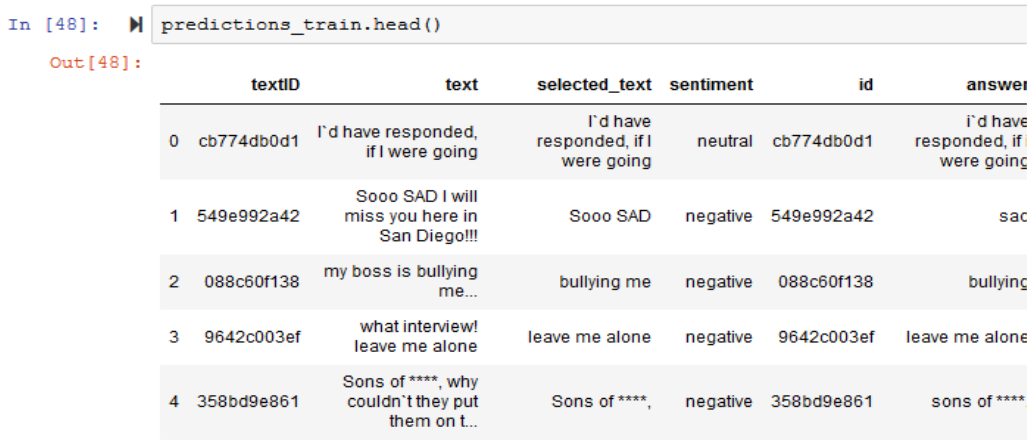
Deep learning approaches have good chance of improving the performance

Transformer based model, DistilBERT is selected and a pretrained model is trained on current dataset by transfer learning

The problem is cast as a Q&A problem where the answer is a selected excerpt from main text

In this case, sentiment is given as ‘question’ input and model is supposed to predict excerpt from the tweet

This shows best performance of all approaches



```
In [48]:   predictions_train.head()
Out[48]:
```

	textID	text	selected_text	sentiment	id	answer
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral	cb774db0d1	i`d have responded, if i were going
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative	549e992a42	sad
2	088c60f138	my boss is bullying me...	bullying me	negative	088c60f138	bullying
3	9642c003ef	what interview! leave me alone	leave me alone	negative	9642c003ef	leave me alone
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative	358bd9e861	sons of ****

Figure 7 Sample predictions of distilbert model showing very good match with the selected text

Conclusions

Deep learning based approaches outperform traditional NLP methods

This shows the power of deep learning based approaches to language understanding

This is a demonstration of rapid progress in NLP recently

DistilBERT model, used in this project **is less than a year old** since first published

Conclusions

Approach	Details	Mean Jaccard score
Trivial score	Return entire tweet as prediction	0.589
Bag-of-words	Return complete tweet for ‘neutral’ sentiment or if original tweet contains fewer than 6 words (ver 3)	0.605
Parse-trees	Return a 5 word interval around the best predictive word as predicted selection. (ver 5)	**0.589**
DistilBERT	*distilbert-base-uncased-distilled-squad’*	0.88 * best performance

Thank You