Real-time sentiment analysis of Tweets using Seq2Seq model and Transformer model

Aishwarya R 312216104005
Preethi M 312216104080
Mentor : Dr. Chandrabose Aravindan

SSN College of Engineering, Chennai

22 September 2020

Abstract

This project performs sentiment analysis on tweets pertinent to a particular topic and classifies the responses into 2 classes:

- Positive
- Negative

Traditional approach involves language modelling. This has major drawbacks.

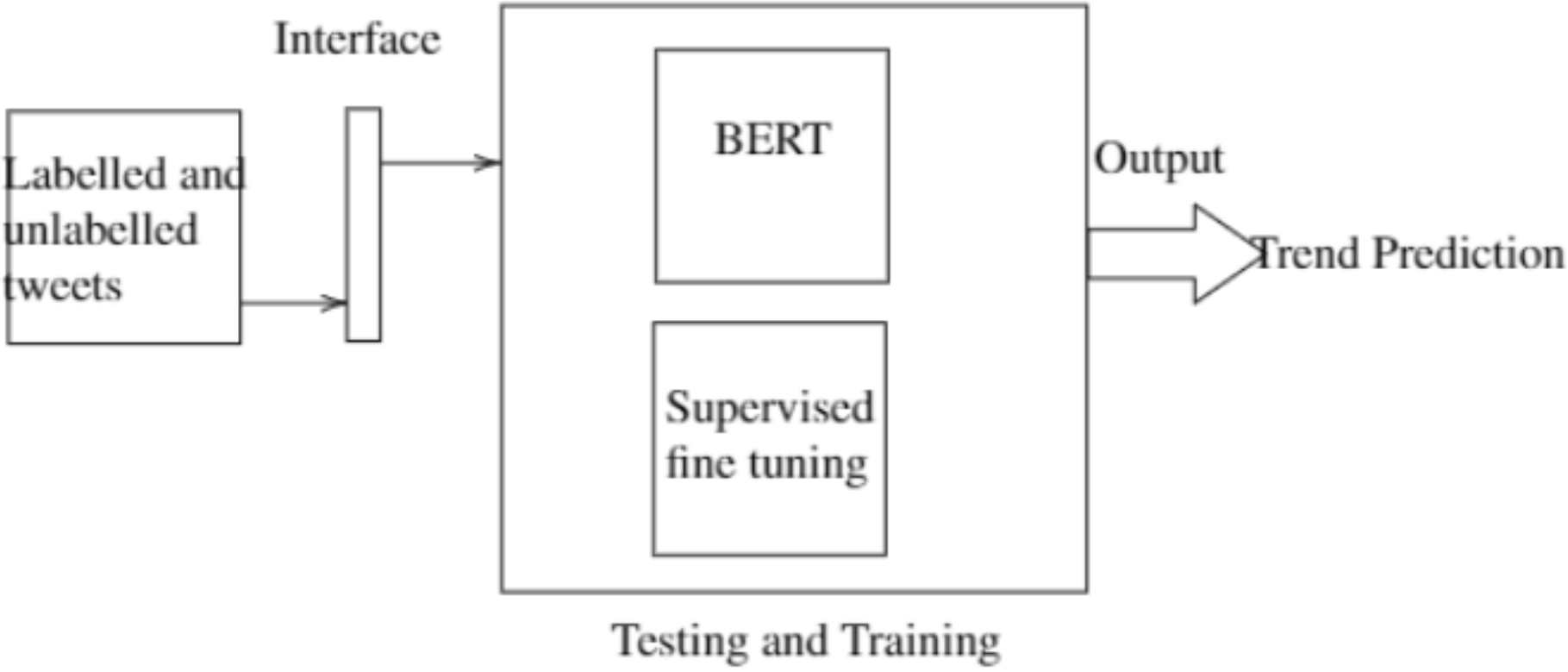
- training a model from scratch requires enormous data, computation resources and is a time consuming process
- trained models usually fail to generalize the scenarios that it has never seen before

Hence Transfer learning is used, where a model trained on one task is re-purposed on a second related task. It is implemented using a transformer model: BERT. A n to 1 mapping Seq2Seq model is built and their performance is compared

Architectural Design For The Proposed System

- The tweets pertinent to a particular topic of interest was first collected from the Twitter Api using Tweepy.
- Expert benchmarking was done with 5 experts.
- The system then uses the model to train and predict the trend seen in majority: positive or negative response.
- Seq2Seq model was built and trained.
- The BERT model is fine-tuned for the task of sentiment analysis of tweets.The model is then cloned.The cloned model is fine-tuned by trained for a very small fraction of the data-set collected for the relevant hashtag.

Architectural Design For The Proposed System



Gold Standard Formulation

Dataset

The sentiment140 is a dataset which contains 1,600,000 tweets extracted using the twitter API . The tweets have been annotated The data is a CSV with emoticons removed. Data file format has 6 fields:

- 0 - the polarity of the tweet (0 = negative,1 = positive)
- 1 - the id of the tweet (2087)
- 2 - the date of the tweet (Sat May 16 23:58:44 UTC 2009)
- 3 - the query (lyx). If there is no query, then this value is NO QUERY.
- 4 - the user that tweeted (robotickilldozr)
- 5 - the text of the tweet (Lyx is cool)

Gold Standard Formulation

Data Acquisition

Tweets pertinent to the following hashtags have been collected for the months from June to September 2019 using Tweepy. After removing retweets, around 800 tweets have been used to form the test corpus.

- PChidambaram
- INXMediaCase
- ChidambaramBail
- SCgrantsbailtoChidambaram
- SupremeCourt
- ChidambaramTimeUp
- TiharJail
- ChidambaramArrested
- ChidambaramFacesJail
- CuriousCaseOfChidambaram

Gold Standard Formulation

Data Preprocessing

Tweepy collects data as objects. Each of these tweet objects have the following fields:

- fullname
- id
- likes
- replies
- retweets
- text
- timestamp

The text field alone required to be extracted. To do this, the text member of the tweet object is aggregated into a pandas data frame. This data frame is stored as a csv by converting to excel. The data from csv file is then imported into google sheets. Further data pre-processing to clean the tweets by removing hashtags, mentions, URLs are while training the respective models.

Gold Standard Formulation

Expert Benchmarking

- Expert bench-marking is carried out by 5 experts. In order to achieve this, the first step is to create a customized form using the tweets stored in the Google Sheets.
- Google Apps script was used to create Google Forms
- The form consists of multiple choice questions. The tweet text is used as questions and the two choices are: positive, negative.

Gold Standard Formulation

- For ease of working, the 750 questions were divided into 8 customized Google Forms and presented to the experts.In this way, the tweets pertaining to a particular topic (here, PC arrest) were presented to 5 human experts to classify the tweets as having positive or negative sentiment.
- The responses from the Form(s) are stored in their corresponding response Sheets.
- Then, they are aggregated into a single Google Sheet.

Gold Standard Formulation

Inter-rater agreement measure

- Experts are called in for advice on their respective subject, but they do not always agree on the particulars of a field of study.
- This is the reason why we calculate the inter-rater agreement. Fleiss's Kappa is an extension of Cohen's kappa for three raters or more.
- Fleiss' kappa, (Fleiss, 1971; Fleiss et al., 2003), is a measure of inter-rater agreement used to determine the level of agreement between two or more raters (also known as "judges" or "observers")
- The kappa value was found to be 0.4061719432.This result indicates a moderate agreement between the different raters

Technologies used

Seq2Seq Model

- A seq2seq bidirectional GRU model is built to perform sentiment analysis.
- It is a n to one model since a sequence of words(tweets) need to be mapped to either positive or negative (sentiment class).
- The validation accuracy was found to be around 80.25 percent
- The model is also tested using the corpus collected as a part of Gold Standard evaluation and the accuracy was found to be 51 percent

Technologies used

Transformer Model

- Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained, neural network-based technique for natural language processing (NLP)
- The BERT model is fine-tuned for the task of sentiment analysis of tweets. It is trained and validated on sentiment140 dataset.
- Hyperparameters such as learning rate have also been tuned for classifying tweets.

Technologies used

Web Application

- A web application has been created in Python. It has been hosted using Streamlit which is an open-source app framework for building Machine Learning and Data Science related data apps.
- The aim of building the app is to make our project available to potential targeted users who would like to know the sentiment response from the people pertinent to a particular hashtag.
- Keeping this in mind, the app consists of a text-box that allows users to enter any hashtag.
- Tweepy is then used to fetch real-time as well as historic tweets based on the hashtag of interest.
- The tweet text is then pre-processed to clean hashtags, mentions. The trained BERT model is loaded. This is then used to perform sentiment classification of the tweets.

Deep Learning Models

Inferences

- Thus, it has been found that the accuracy of the transformer model is better than that of the seq2seq model.
- This is because, transformer models incorporate transfer learning. It focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.

Deep Learning Models

Inferences

Model	Accuracy
Seq2Seq validation accuracy	80.25 %
Seq2Seq prediction accuracy compared to human raters	51%
Transformer validation accuracy	90.5%
Transformer Prediction accuracy compared to human raters	65%
Fine-Tuned Transformer Prediction accuracy compared to human raters	73.53%

Figure: Models and their accuracy

Conclusion

- In this thesis, we have discussed in detail about the importance of sentiment analysis for various organizations to collect useful data.
- Gold Standard corpus has been created using 5 experts and it's importance, factors affecting it have been outlined.
- A web application was built to collect tweets related to a particular topic by searching of tweets using hashtags.
- Seq2Seq model and BERT transformer model have been used to classify the tweets into 2 classes:positive and negative.

Interpretation

- Deep learning consists of artificial neural networks that are modeled on similar networks present in the human brain.
- As data travels through this artificial mesh, each layer processes an aspect of the data, filters outliers, spots familiar entities, and produces the final output.
- Deep learning mimics the human brain.

Interpretation

- In order to do annotation job many organizations have to spend resources like money and time on human raters
- Human raters usually have internal bias and can make errors

Interpretation

- Our web application is open source
- Organizations and policy makers can make use of it as an money saving alternative

FUTURE WORK

- Further scope can be to delve into the usage of a parser that pre-processes the input text by tagging parts of speech.
- Although this is only possible in structured language constructs, it can be experimented with unstructured language constructs such as tweets after cleaning data and expanding contractions to check if accuracy improves.
- Also, the feature of further fine-tuning the trained BERT model specific to desired hashtag by allowing users to manually annotate a small fraction of the tweet text, can be worked on to enhance the functionality of the already developed web application.

Twitter Sentiment Analysis

Search Twitter for Query

Query:

#TajMahal

	tweet	predicted-sentiment
0	RT @moneycontrolcom: As…	1
1	RT @MidasTimes: #Unlock…	1
2	RT @Congress_Army: #Taj…	0
3	RT @RatherNazaket: #Taj…	1
4	RT @sajjanladka: Cases …	1

Positive response

Figure: Web Application

```
Epoch 6/10
 - 13s - loss: 0.0467 - accuracy: 0.9844 - val_loss: 0.8410 - val_accuracy: 0.7825
Epoch 7/10
 - 13s - loss: 0.0397 - accuracy: 0.9875 - val_loss: 0.8560 - val_accuracy: 0.7700
Epoch 8/10
 - 13s - loss: 0.0133 - accuracy: 0.9962 - val_loss: 0.8454 - val_accuracy: 0.8000
Epoch 9/10
 - 13s - loss: 0.0100 - accuracy: 0.9969 - val_loss: 1.2054 - val_accuracy: 0.7650
Epoch 10/10
 - 13s - loss: 0.0128 - accuracy: 0.9975 - val_loss: 0.9838 - val_accuracy: 0.8025
```

Figure: Seq2Seq Model training and validation

	A	B	C	D	E
746	#Chidamb #Chidamb ArnaB kne	0	1	0	
747	Kingpin in just asking #Chidamb	0	0	1	
748	Chidamba Under me Judge : Ok Chidu: Un Judge: Ok Chidu : So J: Ok ,You #Chidamb	0	1	0	
749	Right now Any guess who will la #Chidamb	0	0	1	
750	Opposition	0	1	0	
751	I thought s	0	0	1	
752	There are #Chidamb	0	1	0	
753	#Chidamb	0	0	1	
754					387
755					0.514628
756					

Figure: Seq2Seq Prediction accuracy compared to human raters

```
After training ::
Features loaded from cache at cache_dir/cached_dev_bert_128_2_95
100%|                                                                    | 1
{'mcc': 0.8135345007056454, 'tp': 38, 'tn': 48, 'fp': 7, 'fn': 2, 'acc': 0.9052631578947369, 'eval_loss': 0.2693
    So #Chidambaram goes to Tihar Jail. Judicial custody till 19th September Negative.5
0    P #Chidambaram Likely To Celebrate 74th B'Day ...                           Positive
1    Former PM Manmohan Singh and Congress presiden...                           Negative
2    #Chidambaram Was Terror for women : Dr @swamy3...                           Negative
3    Has N Ram officially joined the Congress? Othe...                           Negative
4    #MustWatchAndShare\nHeart wrenching Narration ...                           Negative
```

Figure: Bert training and validation

	A	B	C	D	E
747	Kingpin in just asking #Chidamb	0	0	1	
748	Chidamba Under me Judge : Ok Chidu: Un Judge: Ok Chidu : So J: Ok ,You #Chidamb	0	1	0	
749	Right now Any guess who will la #Chidamb	0	0	1	
750	Opposition	0	1	0	
751	I thought s	0	0	1	
752	There are #Chidamb	0	0	1	
753	#Chidamb	0	1	0	
754					492
755					0.654255

Figure: Bert Prediction accuracy compared to human raters

Opposition	0	1	0	
I thought s	0	0	1	
There are a handful of Delhi journalists, less than five to be precise, who will be heartbroke n today. Regardless of what the courts decide about the fmr FM, these journalists are guilty of selling their souls. #Chidamb aramFacesJ ail	0	0	1	
#Chidamb	0	1	0	
				553
				0.7353723

Figure: Fine-tuned BERT accuracy compared to human raters

$$p_a = \bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} x_{ij}^2 - m}{m(m-1)} = \frac{1}{mn(m-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2 - mn\right]$$

$$p_e = \sum_{j=1}^{k} q_j^2$$

where

$$q_j = \frac{1}{nm}\sum_{i=1}^{n} x_{ij}$$

Fleiss' Kappa is defined to be

$$\kappa = \frac{p_a - p_e}{1 - p_e}$$

fx =(E4-E5)/(1-E5)

	A	B	C	D	E	F	G	H
1		Positve	Negative					
2		0	5	m	5	no of experts		
3		4	1	n	753	no of records		
4		2	3	pa	0.747875166	(SUMSQ(B2:C754)-E2*E3)/(E2*E3*(E2-1))		
5		0	5	pe	0.5754245171	SUMSQ(B756:C756)		
6		1	4	kappa	0.4061719432	(E4-E5)/(1-E5)		
7		0	5					
8		1	4					
9		1	4					
10		0	5					
11		2	3					
12		5	0					
13		1	4					

fx =1-SUMPRODUCT(B2:B754,E2-B2:B754)/(E2*E3*(E2-1)*B756*(1-B756))

	A	B	C	D	E	F	G	H	I	J	K
748		0	5			negative					
749		2	3		q	SUM(C2:C754)/(E2*E3)					
750		0	5		b	C756*(1-C756)					
751		2	3		k	1-SUMPRODUCT(C2:C754,E2-C2:C754)/(E2*E3*(E2-1)*C756*(1-C756))					
752		0	5		s.e	SQRT(2/(E2*(E2-1)*E3))					
753		0	5		z	C758/C759					
754		0	5								
755						positive					
756	q	0.2932270916	0.6996015936		q	SUM(B2:B754)/(E2*E3)					
757	b	0.2072449644	0.2101592038		b	B756*(1-B756)					
758	k	0.4194371599	0.4154813911		k	1-SUMPRODUCT(B2:B754,E2-B2:B754)/(E2*E3*(E2-1)*B756*(1-B756))					
759	s.e	0.01152398042	0.01152398042		s.e	SQRT(2/(E2*(E2-1)*E3))					
760	z	36.39689972	36.05363561		z	B758/B759					

References


Pennington, J., Socher, R., Salimans, T. and Manning, C., 2014. *GloVe: Global Vectors for Word Representation.*,URL https://www.aclweb.org/anthology/D14-1162.pdf.


Mikolov, T., Chen, K., Corrado , G. and Dean , J., 2013. *Efficient Estimation of Word Representations in Vector Space.*,URL arxiv.org/pdf/1301.3781.pdf.


Ma, X. and Hovey, E., 2016. *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF.*,URL https://arxiv.org/pdf/1603.01354


Lample,G., Ballesteros,M., Subramanian,S., Kawakami,K. and Dyer,C., 2016.*Neural Architectures for Named Entity Recognition* .,URL https://arxiv.org/pdf/1603.01360


Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018.*Deep contextualized word representations.*URL https://arxiv.org/pdf/1802.05365.pdf

Thank You