TARGET BASED STANCE DETECTION

CS 529 PROJECT JAN-MAY, 2022, REPORT

BUDDI KIRAN CHAITANYA, 214161002
IIT GUWAHATI

TOPIC: Target Based Stance Detection

INTRODUCTION:

In this day & age of free speech, while facts are hard to come by, there has been no shortage of opinionated data. The imperative to classify, group & analyse these opinions in an attempt to gauge has led to the rise in popularity of the sentiment analysis or stance detection task within the domain of NLP. Stance detection has implications for democracy, national security, financial market stability, public order, product development, business management etc.

Target-Based Stance Detection (TBSD) or Aspect-based sentiment Analysis (ABSA) aims at unlocking the stance, in favour of, against, or neutral, being expressed, if at all, in the text, towards the queried target/proposition, which at times might not be explicit in the text.

The objective in this phase of the project is to explore the potential of solving the TBSD task using multi-head attention modules that have been fed with contextual word representations (from the pre-trained language model BERT) of the target and the text.

BACKGROUND:

Conventional sentiment classification, deals with classifying the overall sentiment of a text, opinion or a document. Such an aggregate approach often fails to capture the nuance/granularity contained within the opinion and thus might miss out on important information such as the target/entity/topic/aspect towards which the stance is directed at. Stance detection can often go beyond sentiment analysis, because we are seeking the author's subjective outlook towards specific targets/propositions rather than simply about wanting to find out whether the author was happy or angry.

The target may be as varied as a person, a community, an organisation, a government policy, a movement, a product, political events etc. Given a tweet/comment/post and a target entity, the task is to infer whether the author of the text is in favour of the given target, against the given target, or whether neither inference is likely.

Consider the sample target–tweet pairs:

Target: Climate Change is a Real Concern
Tweet: When the Last Tree Is Cut Down, the Last Fish Eaten & the Last Stream Poisoned, You Will Realise That You Cannot Eat Money

Target: Climate Change is a Real Concern
Tweet: Explain the definition of climate change.. #fraud #CCOT #liberty #CruzCrew

Target: Climate Change is a Real Concern
Tweet: I Like learning in depth about nutrition/health @TEDTalks #Disease

The latent stance in the above samples can be inferred as one of 'in favour of' in the first one, 'against' in second sample and 'none' in the last one.

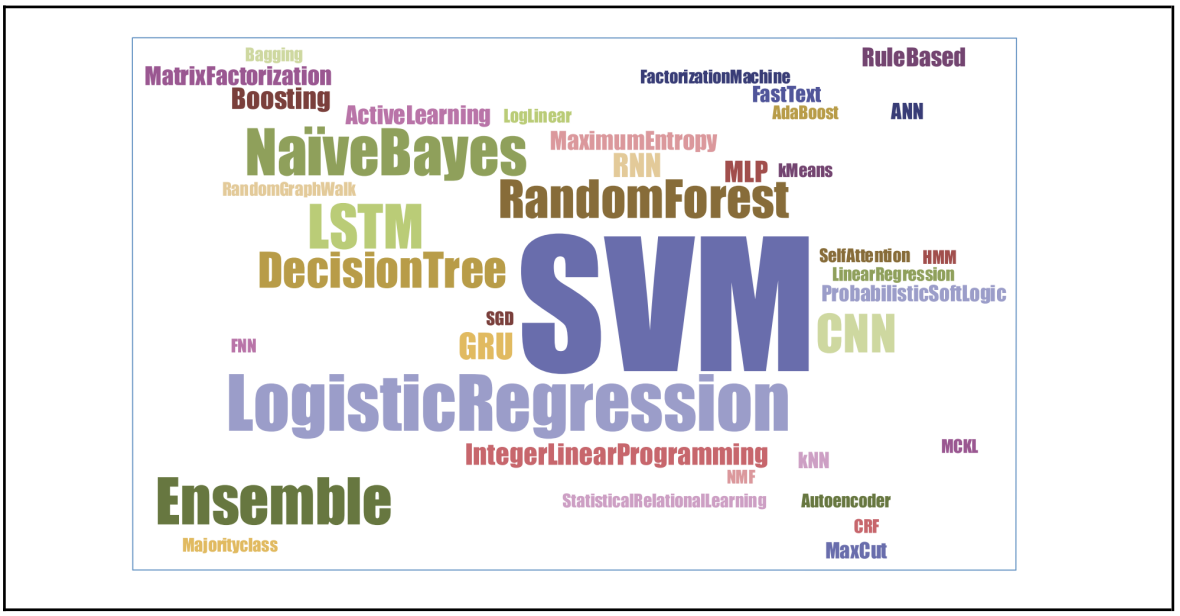
In case the target in the above samples is changed to 'Climate Change is a hoax', the stances with polarity get flipped. To successfully infer the latent stance, an automated model has to identify relevant bits of information that may not be present in the focus text. For example, if one believes that Climate Change is a real concern then s/he is likely against the cutting of trees.

Thus any model that seeks to unlock the latent stance in the tweet, must gain a semantic understanding of the target words/phrase and then use that understanding to probe the text/tweet.

One of the major challenge in targeted stance detection in the context of texts such as tweets is that their brevity often does not allow for the explicit reference to the target word and tweets are often situated within the context of a series of tweets by the user/author. This has issues not only for annotation when tweets are being evaluated at an atomic level but also for stance detection at testing time with unseen targets. There runs the risk of mis-annotating and mis-classification when the text contains little or no reference to the target, even in the embedding space.
An additional issue with twitter is that tweets are informal, full of misspellings, shortenings, and slang. Thus, while the task of stance detection from tweets has clear overlap with related tasks such as argument mining, sentiment analysis, and textual entailment, it exists within a distinct subspace of its own.

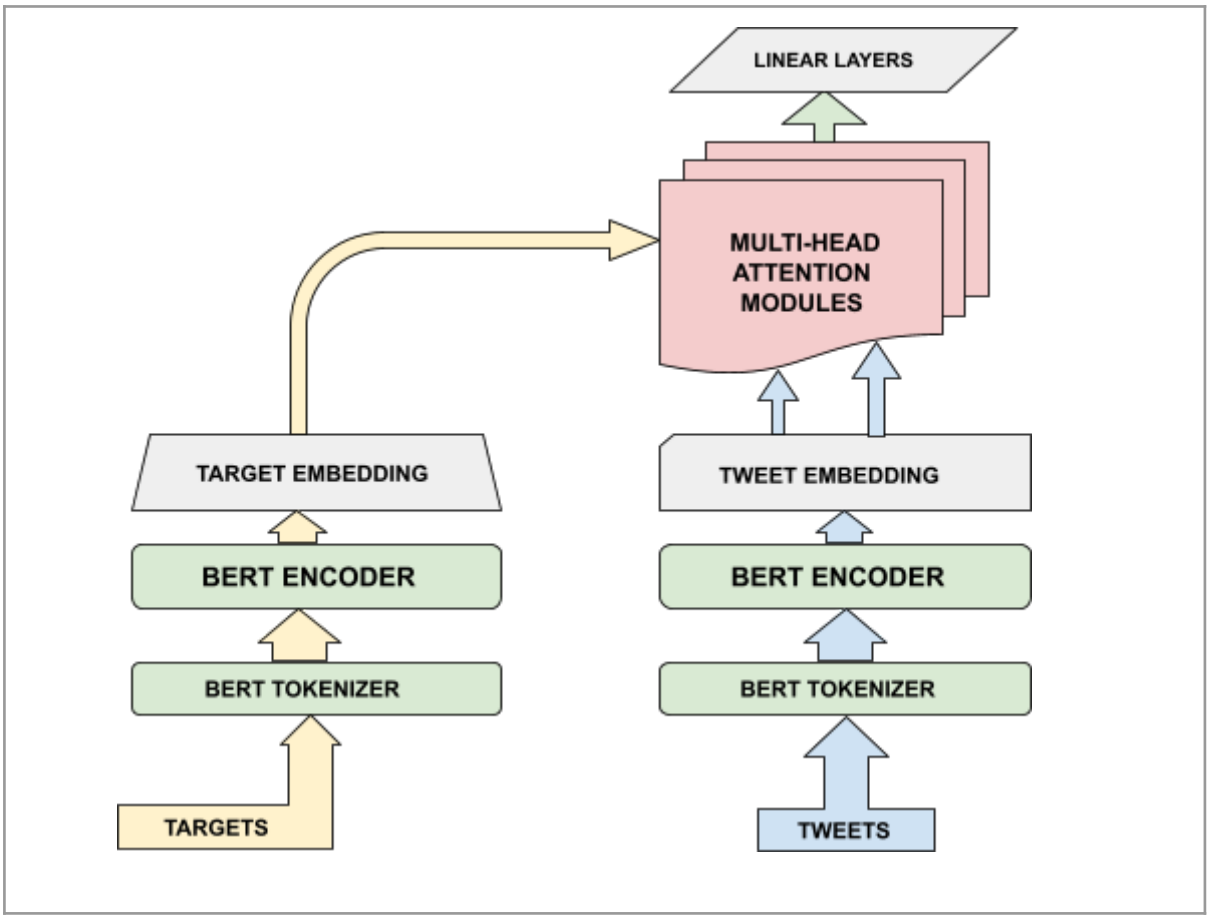
EXISTING APPROACHES TO STANCE DETECTION:

The following word cloud summarises the algorithms that have been typically used to tackle stance detection (pre-2020)

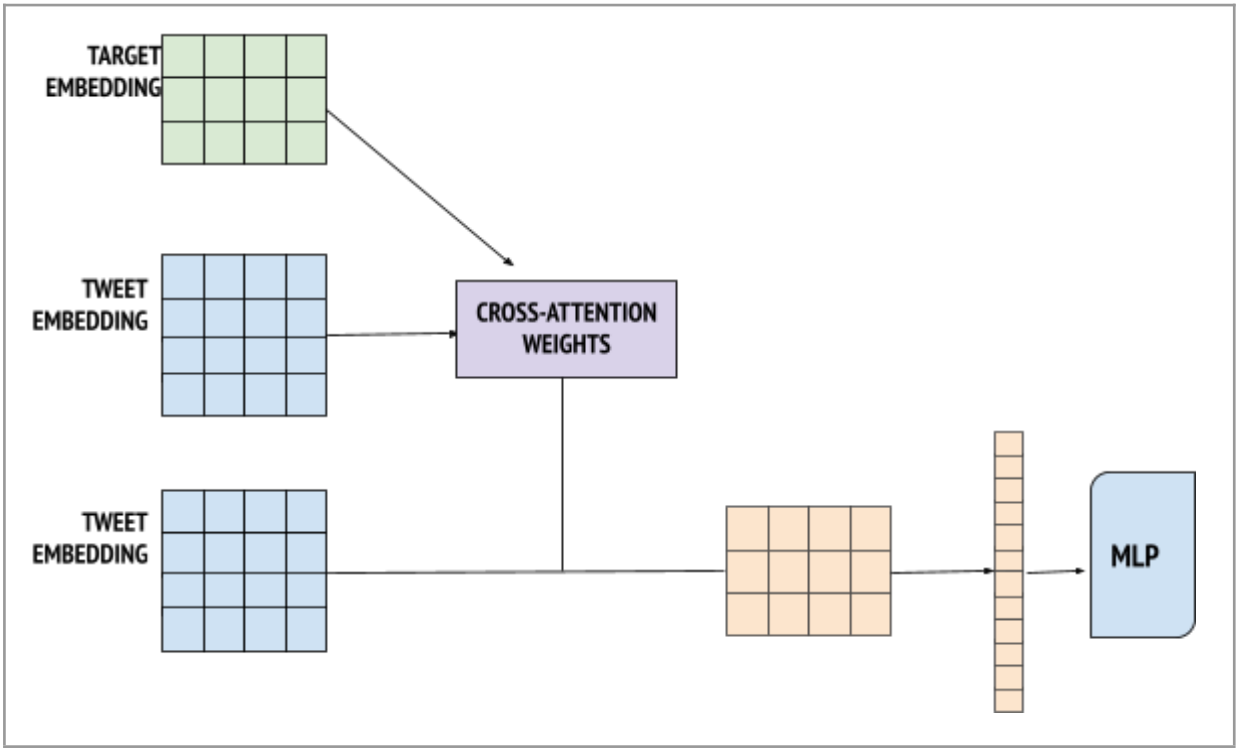


CREDIT: DILEK KÜÇÜK, FAZLI CAN: Stance Detection: A Survey 2020

PROPOSED MODEL:



Our proposed model aims at leveraging the positional & contextual embeddings from a pre-trained model language model (BERT in this case) to reformulate the targeted stance detection to one of cross-attending to the text in question using the target as a query. The embeddings of the target(s) and the tweet/text are fed into a series of multi-head attention modules as query and key/value respectively. The attention modules are then succeeded by linear layers meant for classification. When the weights are fine-tuned, it is hoped that the attention modules would have learnt to focus, utilising the key provided, on those aspects of the text that are of utility in classifying the stance.



To realise the proposed model we have utilised the pre-trained BERT (base-uncased) tokenizer and encoder available via the hugging face repository. Two multi-head attention modules, each with 3 heads, have been realised using the pytorch modules.

The number of attention modules and the heads per module are hyper-parameters. The output from the first attention module is then fed into the other, with the query remaining the same. The final output at the end of the attention modules is then flattened and fed into linear/classification layers (2 hidden layers and 1 output layer). Cross Entropy loss function along with the optimizer Adam and a batch size of 128, were utilised during training.

Another set of hyper parameters are the maximum length of the targets and tweets that we desire, which in our model have been set to 5 and 50 respectively, with appropriate padding/truncation used to achieve the same.

DATASET: SemEval-2016 Task 6: Detecting Stance in Tweets

A dataset of 2914 manually annotated tweets for stance towards given target, target of opinion (opinion towards), and sentiment (polarity). Each sample/row in the dataset consists of a tweet, a target entity (person, organisation, etc.), and an annotated stance. The target of interest may or may not be referred to in the tweet, and it may or may not be the target of opinion.

We have decided to tackle the Task A of this challenge, where a part (80%) of the dataset has been used to train the model, and the results evaluated upon the remaining hold-out set. The implication being that we do not encounter any novel targets during the testing phase. But we do not see any reason for the model not generalising well to novel targets for we are operating in the embedding space where a target of 'Climate Change' would be situated close to a target such as 'Environmental Protection', thus resulting in the similar cross-attention outputs and eventually similar stance classification.

The possible candidates for targets in this dataset are:

- Atheism
- Climate Change is a Real Concern
- Feminist Movement
- Hillary Clinton
- Legalization of Abortion

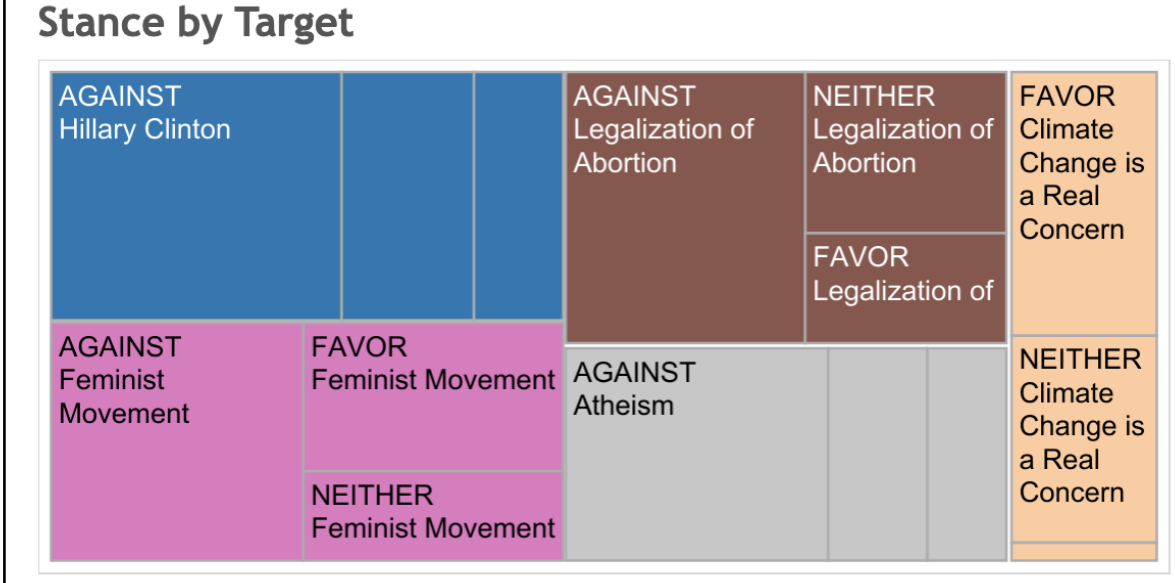
The possible stance categories in this dataset are:

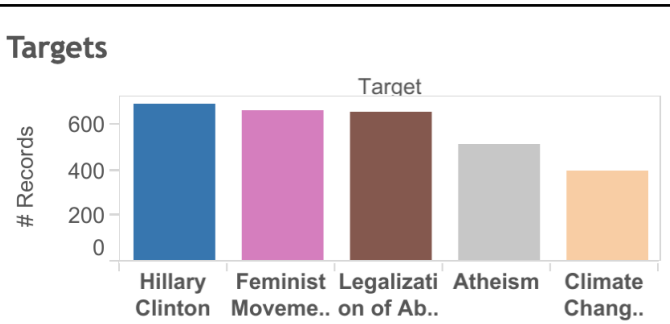
- FAVOUR: we can infer that the tweeter supports the target (e.g., directly or indirectly by supporting someone/something, by opposing or criticising someone/something opposed to the target, or by echoing the stance of somebody else).
- AGAINST: we can infer that the tweeter is against the target (e.g., directly or indirectly by opposing or criticising someone/something, by supporting someone/something opposed to the target, or by echoing the stance of somebody else).

- NONE: none of such above inference possible.

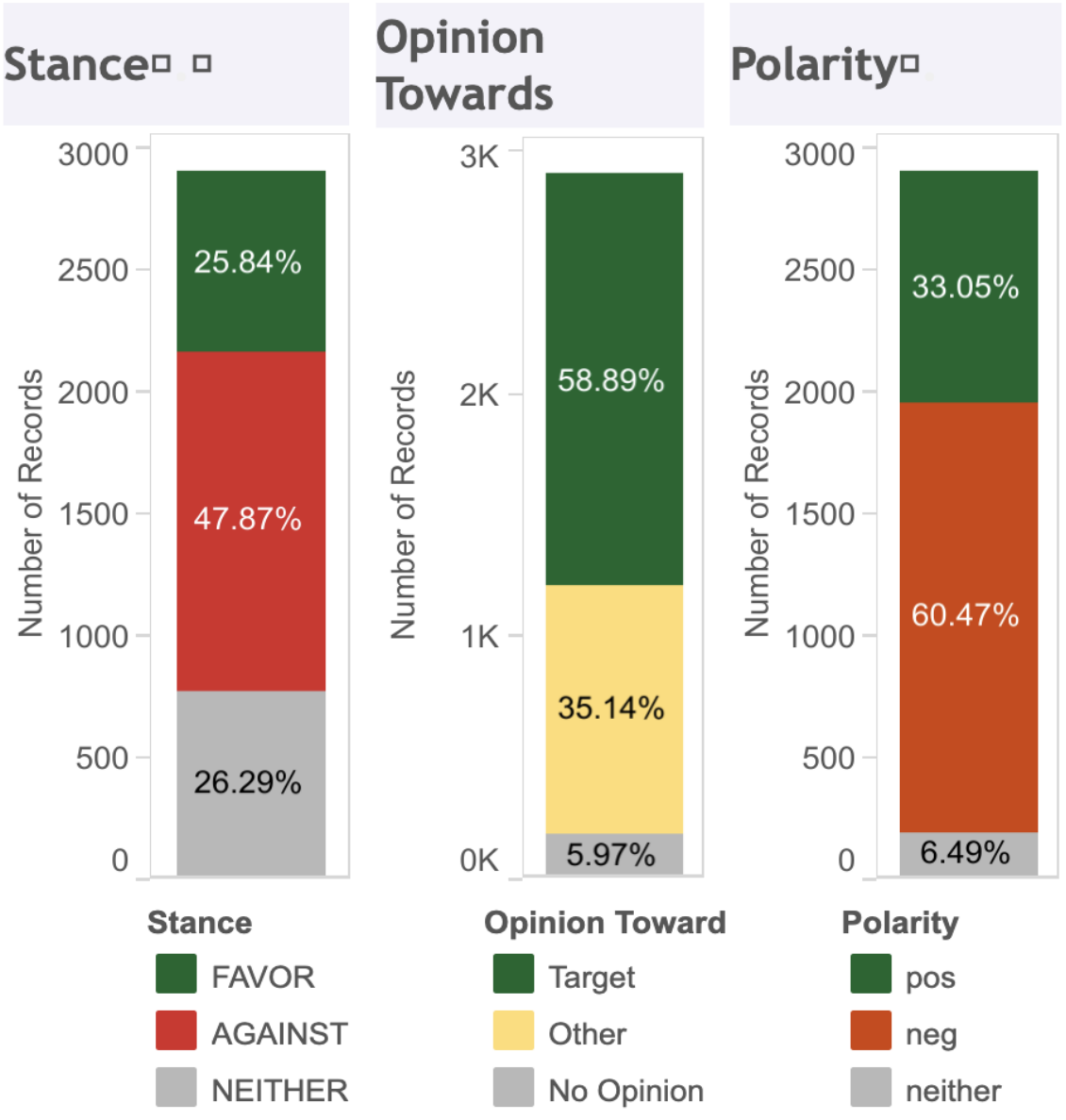
The following visualisations break down the dataset in terms of the distribution over targets and the classes:

STANCES BY TARGET





BREAKDOWN BY STANCE CLASS, POLARITY AND OPINION TOWARDS



X by Y Matrices

	Opinion Toward		
Stance	Target	Other	No Opinio..
FAVOR	93.89%	5.18%	0.93%
AGAINST	71.68%	27.74%	0.57%
NEITHER	1.17%	78.07%	20.76%

	Sentiment labels		
Stanc	pos	neg	neither
FAVOR	35.99%	55.78%	8.23%
AGAINST	30.68%	66.74%	2.58%
NEITHER	34.46%	53.66%	11.88%

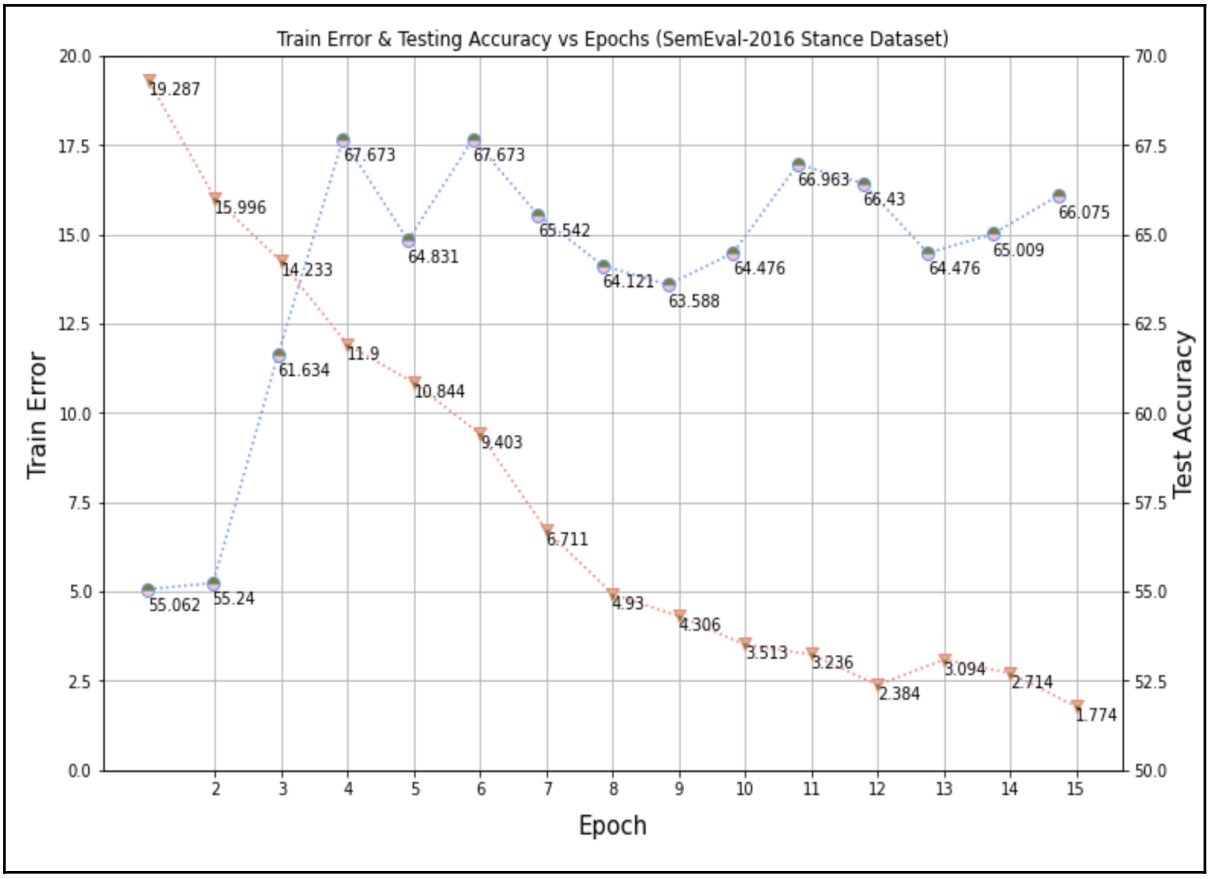
	Sentiment labels		
Opinion To..	pos	neg	neither
Target	29.08%	66.43%	4.49%
Other	38.28%	55.37%	6.35%
No Opinion	41.38%	31.61%	27.01%

In the original competition that accompanied the debut of this dataset in 2016, participants have employed traditional feature-based machine learning, deep learning, and combined (ensemble) methods. The best performing system for subtask A was a RNN-based system that had attained an F-score of 67.82%, while a system based on CNNs ranked second for subtask A with an F-score of 67.33%. The baseline system using an SVM-based approach provided by the shared task organisers attains an F-score of 68.98% for subtask A.

EXPERIMENTS & RESULTS:

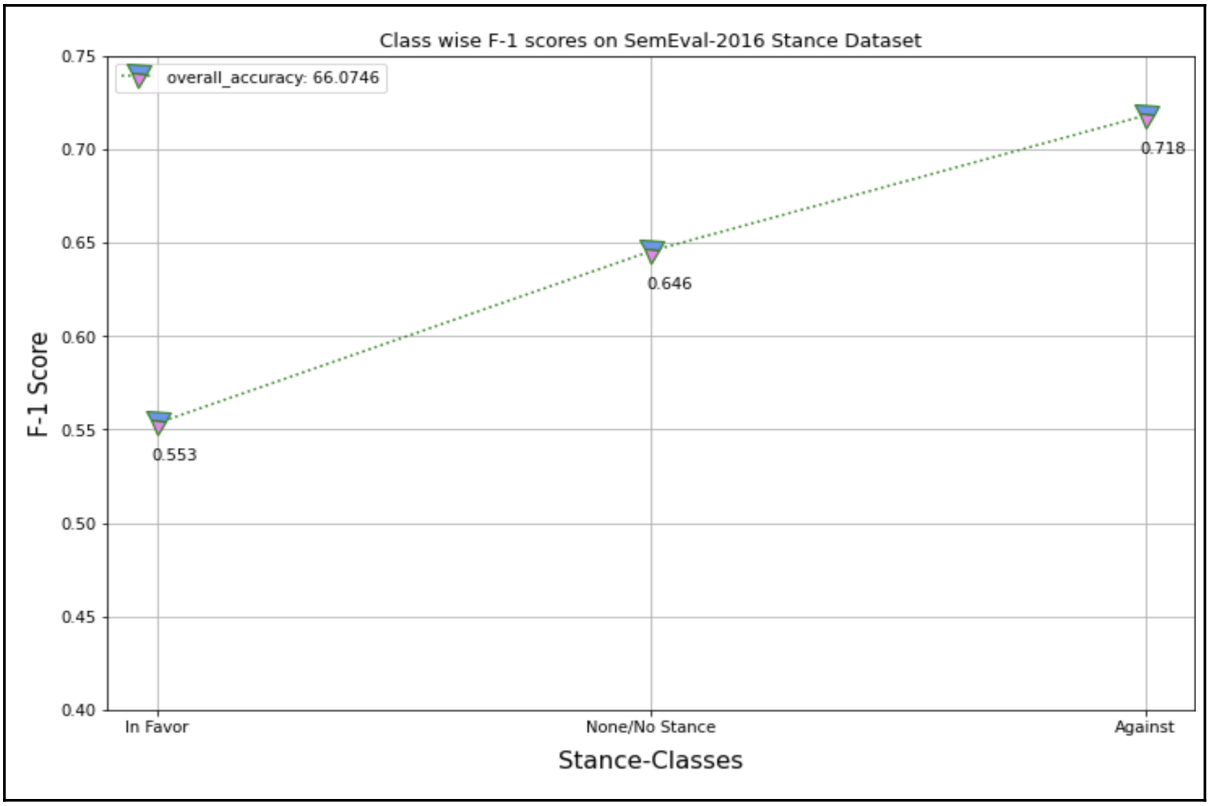
The training error and testing accuracy over epochs are tabulated and visualised below

EPOCH	TRAIN ERROR	TEST ACCURACY
EPOCH 1	19.2867	55.0622
EPOCH 2	15.996	55.2398
EPOCH 3	14.2327	61.6341
EPOCH 4	11.9003	67.6732
EPOCH 5	10.8443	64.8313
EPOCH 6	9.40329	67.6732
EPOCH 7	6.71081	65.5417
EPOCH 8	4.92979	64.1208
EPOCH 9	4.30583	63.5879
EPOCH 10	3.51288	64.476
EPOCH 11	3.23633	66.9627
EPOCH 12	2.38356	66.4298
EPOCH 13	3.09363	64.476
EPOCH 14	2.71442	65.0089
EPOCH 15	1.7735	66.0746

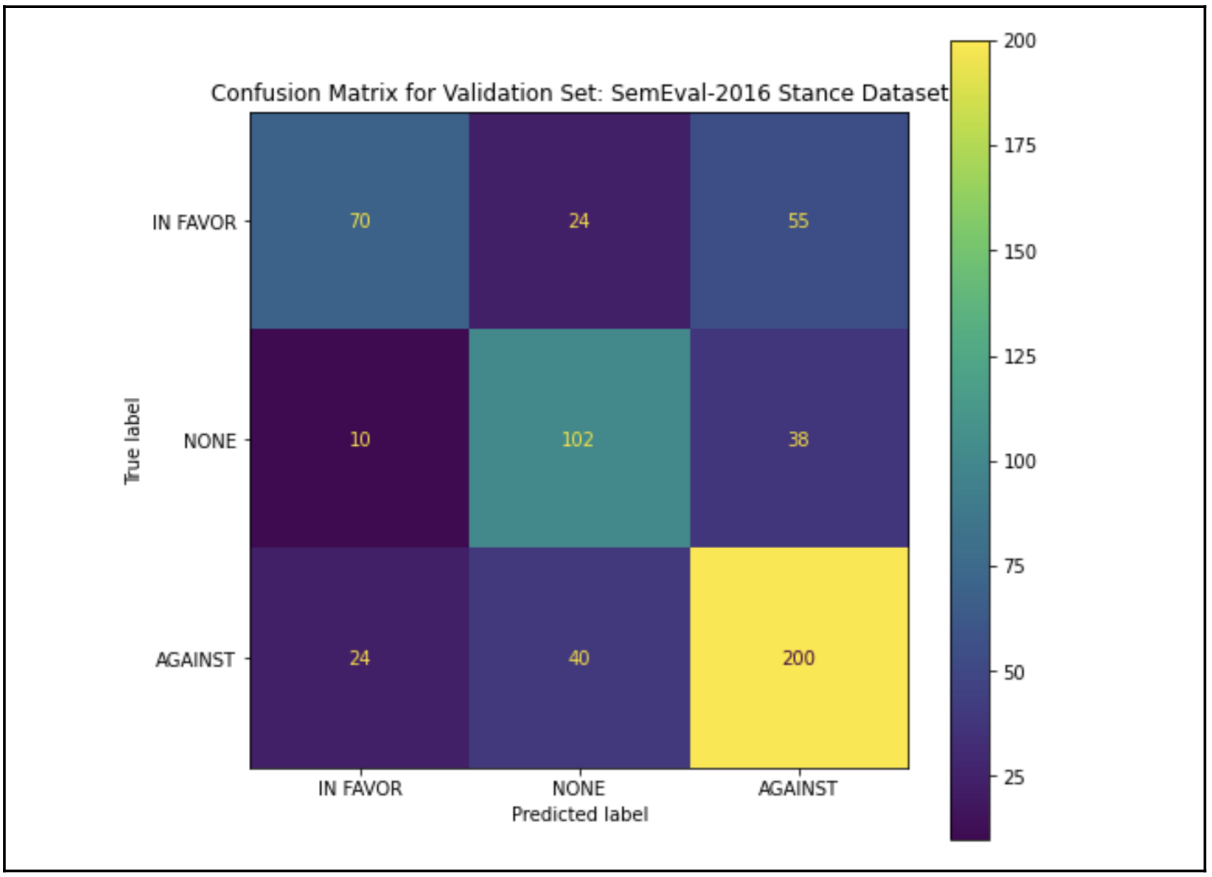


CLASS WISE F1-SCORES

ACCURACY	F1 (CLASS: IN FAVOR)	F1 (CLASS: NONE)	F1 (CLASS: AGAINST)
0.660746	0.55336	0.64557	0.718133



CONFUSION MATRIX



FUTURE WORK/SCOPE:

- Exploring the possibility of extracting features from hashtags that could aid stance detection. Hashtags in the current model were treated as any other words, which is not optimal considering they contain more information per word in relation to the stance being expressed.

- Exploring the possibility of mining the original tweet, in case of retweets & forwards, to gauge the stance of the retweeter when no explicit stance is being expressed in the retweet.

- Testing the model on a much larger data set with unseen targets, and

- Algorithmic optimization to handle the BERT embeddings in a less computationally intensive manner.

REFERENCES:

Mohammed M.Abdelgwada, etal.: Arabic aspect based sentiment classification using BERT
Mickel Hoang, Oskar Alija Bihorac, Jacobo Rouces: Aspect-Based Sentiment Analysis Using BERT
Saif M. Mohammad, Svetlana Kiritchenko: SemEval-2016 Task 6: Detecting Stance in Tweets
DILEK KÜÇÜK, FAZLI CAN: Stance Detection: A Survey