Stance Classification in Tweets

1. Abstract

The main objective of this report is showcase the study behind stance classification of tweet and discovering the best fitting deep learning model that can successfully classify stance (FAVOR, AGAINST, NONE) based on particular Target (topic) about which the tweet is written about. After intense research on various models, Final selection was made of LSTM model as they proved to work best in natural language processing scenario. Analysing of intra variability of classes and imbalancing of data, efforts were taken choose a balanced data set for training purpose but deep learning model demands large data thus transfer learning methodology was used to help solve this. Using knowledge from model glove trained trained on wikipedia data which is also a source of informal vocabulary like twitter. The data was transformed to task vocabulary to fit model. The model after large iterations of fine tuning making it robust and provided an accuracy score on validation as 52.97 and test score as 53.8 respectively. The possible ways to get better accuracy woulda be using sentiment analyses data for categorisation and training each target tweets seperate as vocabulary similarity will be better and scrapping data to increase training dataset and also transfer learning from model run on twitter data will be ideal.

2. Introduction

Stance classification is basically interpreting from a given text whether the author is in Favour, Against or is Neutral towards the proposition or target mentioned in the text. This task is even complex for human minds since it requires factors like reasoning, subjectivity, knowledge and sentiments. Consider the target - tweet pair

Target : Legalization of Abortion
Tweet :You were alive for up to 9 months before you were bornKills #SemST

Humans can infer from the above pair that the author is most likely against the target topic mentioned. The targets can range from a person, product, movement etc. Here we use twitter data which is much less structured and dialogical in nature due to use of non-standard language. The models need to be trained to follow up this literature for it to identify the stance of the tweeter on the target. Automatically detecting stance has widespread applications in fields like information retrieval, textual entailment and text summarisation [1]. Detection of stance in political arenas is the most researched field in this genre [2].

3. Dataset

The data set given is a multi-class dataset, consisting of a training set of data and testing set of data. The training set contains 2,814 set of tweets and testing with 1,249 tweets, and its corresponding labelled information on the target towards which the tweeter shows a stance, also the Opinion towards and sentiment classes are also present. The annotated stage are Favor, Against or None.The five targets we will handle in the project include Atheism, Climate change is a Real Concern, Feminist Movement, Hilary Clinton and Legalization of Abortion. The original dataset is from “Semeval-2016 Task 6: Detecting Stance in Tweets [3].

Concern 1 : The size of the training set is a major concern since its in the nature of deep learning that it functions extensively only on large dataset, but 1,249 tweets divided into 5 targets should lead to inadequate training on models.

Concern 2 : imbalanced dataset is a major concern detected in data exploration phase where the data was divided via target and a bar plot as shown in Fig. 1, revealed concerning low level of tweets available for “Climate change is a Real Concern” which was approximately half in size as compared to other four targets which between them seemed fairly balanced ._This variability could

lead to improper training of model which could lead to improper classification in “Climate change is a Real Concern” category. Further creation of mosaicplot as shown in Fig. 2, showed concerns of biased distribution of stance in each target , with against stance taking higher precedence in all target categories other than “Climate change is a Real Concern” where it had alarmingly very low tweets available in stance - Against (less that 4%). Thus imbalancing is present both within and between each class.

Concern 3 : Random pooling of tweets from each target ,showed presence of text data which is specific to twitter linguistic but will not help us in classification purpose , for example mentions “@Britney, @nomadicvoice” , emoji’s (;), :) , :-)))) , :-)) , punctuation (-, …,/) which urged the need of extensive data cleaning.

Concern 4 : presence of target in tweet provided means for target-dependent classification, but further analysis on original data provided

Target : Donald Trump

Tweet :My vote is definitely for Hillary. Can’t trust #gop candidates

This ruled out target-dependent classification as target is Donald Trump but tweet shows Hillary in its sentence.

4. Methodology

Twitter data is hotbed for stance detection and sentiment analysis in recent times mainly because in twitter the author can express a stance in unstructured, informal or incoherent manner, mostly in a single line and has a 140 character limit if each tweet. It is closely knit with sentiment analysis, since stance analysis often bring information pertaining to sentiment analysis, because in stance, we also look into the tweeters evaluative outlook to the specific targets for example being neutral towards the target, rather than just considering the authors happiness or anger [4]. However there is a significant difference between stance and sentiment. In sentiment we fix upon if a given text is positive, negative or neutral or we determine the speakers opinion and the corresponding target the opinion is expressed towards whereas in Stance we determine favourability towards a target which is already pre determined. But target may not be necessarily mentioned in the text.

A seminal work by Mohammad, Sobhani and Kiritchenko. [6], followed by a SemEval 2016 task [7] conducted by the authors, resulted in kicking-off extensive research in this field. Multiple dimensions of the target tweet combination is explored to identify various phenomena like presence of target word in tweet or presence of any other target word in tweet but rightful owner of tweet being some other target. the fact that if there is not much evidence the tweet is favour or against the target does not imply its automatically neutral, this makes things increasingly tough because mostly a author is inspired to tweet on a target based on a favourable or unfavourable notion in regards to target. The solution model suggested here is the highest accuracy model among all the attempts in stance classification a F-score of 70.3 was recorded which was calculated by averaging the F-favor and F-against, the model used was a linear-kernel SVM classifier as SVMs have proven to be effective on text categorization. This works on features like words and n-grams ,drawn from the training instances. Other than this it works on features obtained using external resources like word embedding which is done on heaps or corpus which is unlabelled. As part of knowledge understanding various experiments were done which provided evidence that stance detection in tweets which express opinion on entity other than the target is hard to detect.

The logic behind data creation was using hashtags related to target as search keywords to find appropriate match in the twitter database which was connected using the Twitter API, to elaborate, a pool of hashtags was created in with three categories as shown below.

Favor Hashtags	Against Hashtags	Stance-ambiguous Hashtags
#Hillary4President	*#HillNo*	*#Hillary2016*

Another approach could be use of emoticons which could help fetch data by mapping :) with positive and :(with negative. Using a strong n-gram resulted in results better than models that used deep learning methods, mainly it could be because of lack of performance of deep learning models due to smaller dataset. Use of Word2Vec as word embedding gave comparatively lesser

value for Hillary Clinton target. Use of sentiment features provided for better stance detection but in our project we need to focus on classifying stance into favour, against and neutral categories.

To make the model learn more of the twitter linguistic we can use the English POS tagger that is designed especially for Twitter data as mentioned inCMU Twitter NLP tool [8] and is created by manually tagging 1,827 tweets. Use of deep learning methods can produce results with wider dependencies due to its back propagation logic, to implement these methods we need to find a wa to investigate this a model based on CNN was evaluated . To implement this we need to classify the words or sentences, Yoon Kim on his paper on CNN for sentence classification [9] provides a method by which a convolution layer applied on top of a CNN model in which first the vectors are obtained via running them on huge corpus of Google news data using Mikolov method . Now using this we implement a model where word embedding is done from Google News database with 300 dimensionality. The distinguishing feature of the modelling was use of vote scheme to predict test set labels using results from softmax [10] .Cross validation is implemented by taking ten parallel epochs, whose validation sets are randomly selected from the training data. Prediction is usually done when our model provides a good accuracy but here some iterations are deliberately chosen to predict the test set and in each iteration when the epoch ends a label is chosen which appears most commonly in the predictions.The convolution layer used here is targeted to detect patterns in tweets which is done using filter matrix which has same rows as input sentence matrix and slides over the columns to provide a feature map. Other features of the model include inputting. Bias vector element wise and use of ReLU as activation function also drop out is set to 0.5. Pooling is done using max-pooling and softmax is used as output layer which calculates the probability of each class to choose predicted label having maximum value. During modelling the training dataset is divided into five based on given targets and each sub division is modelled separately since there is a good chance words to be more similar to each other when talking about same target rather than as a single corpus unit.

Another approach to mitigate these issues was focused in a model proposed by Guid and Amy in there paper [12] where they used RNN network with four layers of weights as shown in Fig 3 and word2vec skip-gram method for word embedding also the model trained on two large unlabelled dataset via distant supervision and used weights used in this learning to given training dataset. The unlabelled dataset was created via sampling 218,179,858 tweets using twitter api. The data was cleaned and transformed to mimic the training dataset. Then used word2phrase to detect phrases containing up to four words. The distinguishing feature of their modelling was using features for learning sentences present in training data via a hashtag prediction. 197 hashtag were queried using nearest-neighbor search and used these hashtags to extract 298,973 tweets containing at least one of these hashtags. Once these sentence vectors were set they were then used for stance detection. The modelling configuration consisted of embedding layer with 256 dimensions using one hot encoding with each token getting a binary vector in the index showing tokens position in the vocabulary. The recurrent layer contains 128 LSTM units which received input as series of up to 30 embeddings and had weights initialised from training on unlabelled data from transfer learning mechanism so that the network learned distributed sentence representations from sampled unlabelled dataset. The next layer is the Rectified linear units using dropouts this layer is connected to LSTM layer by connecting the terminal output of LSTM to 128 dimension layer of Rectified linear. The output layer is a softmax layer containing three dimensions since our perdition is to either favour, against or none. The network were trained with stochastic gradient descent and used a learning rate of 0.015 and momentum value of 0.9. Training was done using categorical cross-entropy loss function. Epochs used was 50 with early stopping to get optimum validation loss.This approach gave a F1 score of 67.8 but only took favour and against stance into consideration.

Success with LSTM but failure to take in neutral stance into consideration, led to further research to get a Two-Phase LSTM model technique using attention for stance detection provided a F-score of 68.84% and best-case accuracy of 60.2% [5]. This approach took perform accuracy measurements with all three stances (favour, against and none). The idea behind this is that neutral stances are usually non-subjective. Therefore as first phase of the approach LSTM is used to classify into subjective (favor/against) and non subjective (neutral). In second phase again LSTM is used to find sentiment (favor/against) of subjective subcategory from first phase. And additional modelling technique used here comprise of attention modelling where augmentation of word embeddings with target topics is used and this is passed through a linear layer for determining the attention of each word in the tweet and context of the topic under consideration. To calculate attention we average embedding of the target under consideration is

Fig 3

taken and augmented with embedding of the constituent words.Both SGD and Adam optimisers and model is trained using cross-entropy loss function so that loss occurring during the first phase is not passed on to the other phase.

5. Modelling

Exploratory Data Analysis - Preliminary analysis of data raised major concerns as discussed earlier. Further analysis is done to extract features from data. Analysing the stance distribution of each Target type provided major concern of presence of scarce data in none category Analysing the stance distribution of each Target type provided evidence, but surprisingly for Climate change as shown in Fig 4, there were a lot of stances with none(neutral) labels. Another major part of data was presence of hashtags, as part of hashtag analysis all hashtags were collected using “#” keyword and plot of top 10 most frequent hashtags shown in Fig 5, provided presence of #SemST as most common, followed by #WakeUpAmerica and #HillaryClinton. Exploring into SemST as a useful feature failed as in “Stance or insults?” Paper [13] it says some hashtags used as queries to extract tweets were replaced with "#SemST" to exclude obvious cues for the classification.

Data Preparation - To avoid garage in garbage we need to clean the given data. Cleaning of data is a vital step. Rule based learning units like spacey and nltk is used to prepossess the text. Stop words are words like "ok", "hmm", “oh", "yep" etc are of very little use. It is better to get rid of them, we use nltk to remove stop words and lemmatize the word in later part for better meaning that that matches the pre trained glove data on transfer learning. Next set of iterative cleaning included removing mentions identified using “@“ keyword. Mentions are usernames that adds no meaning to target or stance of the tweet hence removing them. Iterative checking of data after each cleaning step provided us with variety of problems shown in table and actions taken. Spacey lemmatizer is used to replace with the vocabulary data and the text data contains ascii values which are ignored in preprocessing.

Error	Reason	Action
Puntuations	Words like !”#$%&'()*+,-./:;<=>_`{|}~ don’t add meaning in classification	Removed
Extra spaces	Presence of double spaces noticed, will effect tokenising	Removed
1 & 2 letter words	Even after removing stop words single & dual character that not helpful	Removed
AlphaNumeric	Presence of junks of alphanumeric character and numbers	Removed
ASCII	presence of ascii value instead of original English word.	

The first part is to create a task vocabulary which includes tokenising the data by which all unique words are extracted from all the tweets and each unique word is given a unique number that in converting it from text to sequence order and on testing on our data we have 6318 unique words. To make sure all sentences have same number of elements, which is essential for creating the matrix, we use a pad sequencing which added zeros to empty space.
To model the data we split the given training data into train and validation with a 70/30 split in random order. Ideally we should split It on basis of target but lack of data in climate change does not facilitate that.On choosing the model the sequential LSTM takes more time to train and it takes two separate model to fit into one as we have two classifiers Target and Stance which make it difficult to tune the model.
Model we choose here is bidirectional LSTM where we can pass the Target and Stance in the bidirectional LSTM and add them in chain with the pre trained glove data on the embedding layer. We don’t need to do one hot encoding because the tenser flow embedding layer takes care of it as when we convert each word with index according to the vocabulary, the embedding layer will look after the encoding and provide efficient inference of the encoding. LSTMs can capture relationship between sequence of words

The Target and tweet are passed through as list and the layer is passed with stance activation as SoftMax and we freeze the weights in the model and use the model to fit the test data. The given text classification as imbalance in Target on Climate Change which affects the model accuracy.Major issue faced was case of overfitting since using higher epoch meant model doesn’t have sufficient data to train with , to solve this issue we used

methods mentioned in by Srivastava in his paper [11], where idea is to randomly drop units with their connections from the neural network during training.The best accuracy can be achieved by using grid search that gives the best tuned values than using the for loop which was used in the model for parameter tuning for best accuracy.The parameters are tuned by changing the neurons in the model and the dropout to avoid overfitting. The parameters which gave good results dropouts (0.5,0.7) and the dense layers are added according to the accuracy as well.We use F1 accuracy score as the validating part and the loss is calculated for overfitting.The accuracy score of final model on validation is 52.97 and test score is 53.8. The Fig 7 shows the plots of model and as clearly seen there is no trace of overfitting and current accuracy received is rightfully trained. The confusion matrix is shown in Fig 8.Confusion matrix through test case result it predicted the 516 tweets correctly as Against and 90 tweets as Favor. Where AGAINST and FAVOR are encoded with 0's and 1's.

Summary of Steps taken in Creating and tuning a model

1)on modelling we create a function and pass the parameters for kernal_regulariser and learning rate to pass on the layers.

2) then we create two lstm models to pass on the unit size,drop out and sequences where the first model is loaded with the weights of the embedding layer.

3)the second model is loaded with the chain of the first lstm model then we use merging to concatenate the target with second model

4)we add the layers with Relu activation what is a function used in deep learning and the we use to tune drop out to handle overfiting.

5)The model uses regularization to avoid overfitting on the data where we use early stop to detect the loss in the model and get the accuracy which reduces the model run time and l2 regularizers is also tuned{0.0001,0.0002,0.0005} .

6. Ultimate Judgement

The model chosen performs well when fine tuned the hyper parameters. The model can be improved with different modelling technique that are available even with the imbalance of the data. Using Transfer learning glove 200d twitter pre trained data the model performs well with the better accuracy. Then the second approach on training the target with tweet and then tuning it with stance will create lot of difficulties in tuning it which take more time.On comparing these two models, bidirectional is one approach where tuning is done for one model the classifiers are passed simultaneously to predict the stance. Final configuration of model used is shown in Fig 9. Taking knowledge from research on field one can obtain better accurancy buy first balancing the dataset, specially for climate change, this can be achieved by scrapping data which was attempted by us but was not successful. Also transferring weights from model trained on twitter data would be more ideal. Handling each target separately is proved to be better.

7. Independent Evaluation

For independent evaluation of the final chosen mode, we created a dataset by manual scrapping of data from twitter using hashtags fetched during hashtag analysis and shown in Fig 6 and human reasoning was made to cross check if the tweet fetched provided right stance and right target. The data set created was a balanced dataset, and was treated with same pre processing steps as used to clean the training data . The accuracy is evaluated using F1 score here and the score was 30.6. The increasingly low value of accuracy compared to training and testing evaluation could be because of low sample data as only 5 tweets per stance for each topic is taken. Also as time passes by peoples informal vocabulary changes and the training data was made using tweets used in 2016 and tweets chosen here for most targets were as latest as possible thus unable to match is a scenario.

APPENDIX

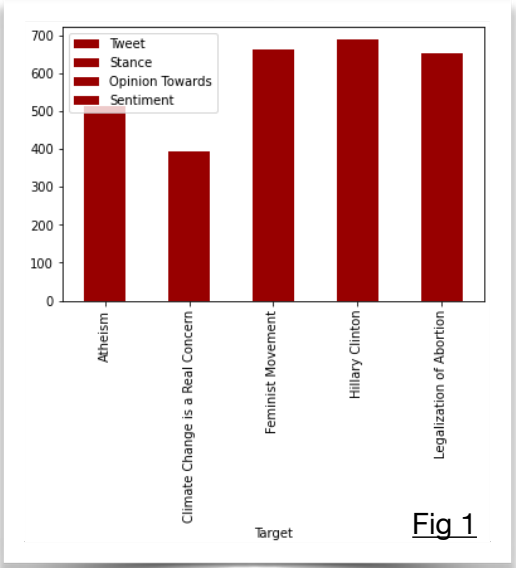


Fig 1

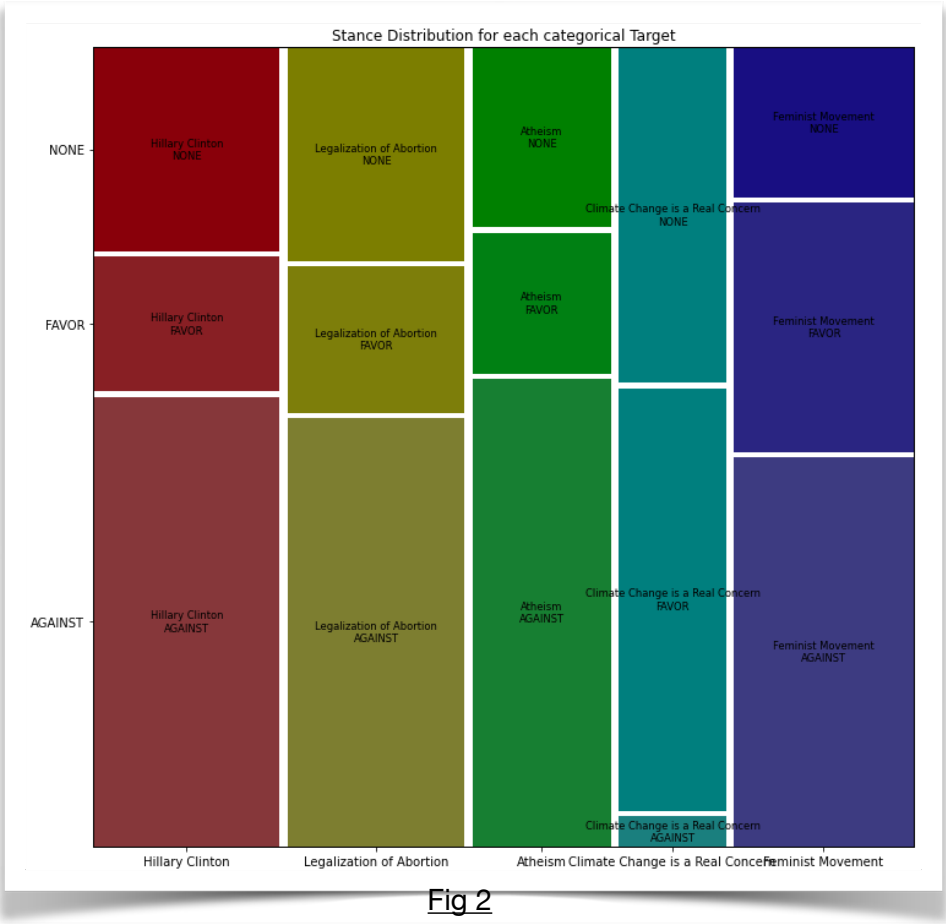


Fig 2

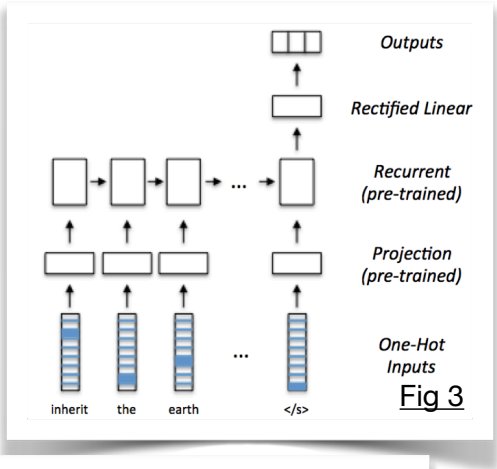


Fig 3

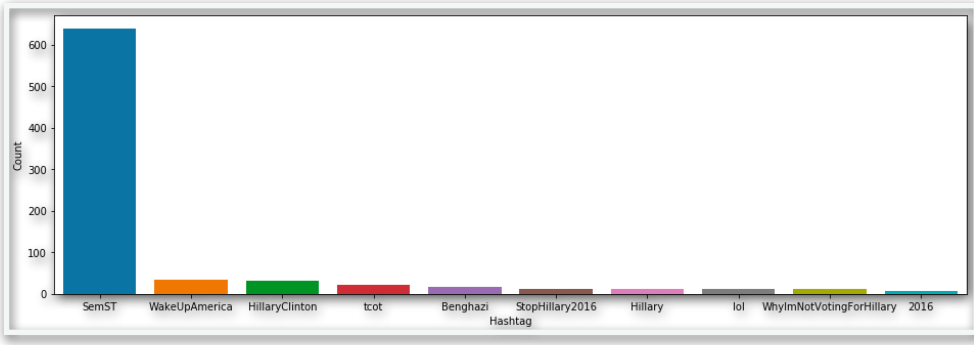


Fig 5

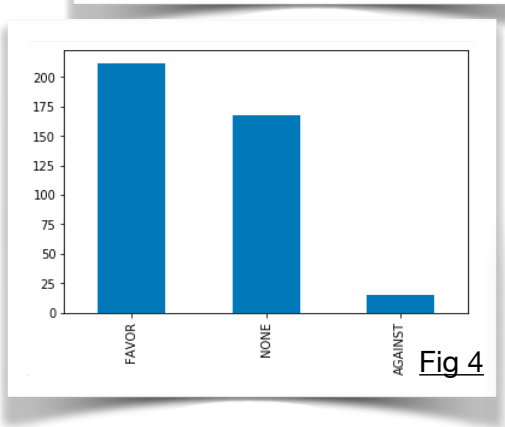
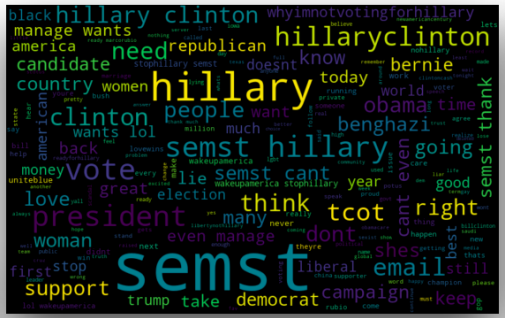
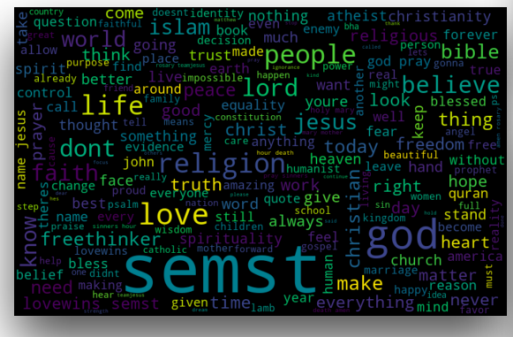


Fig 4

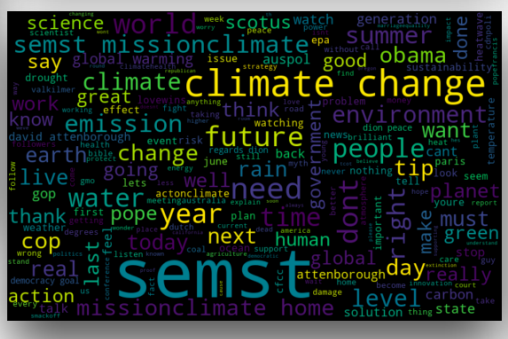


Hillary Clinton

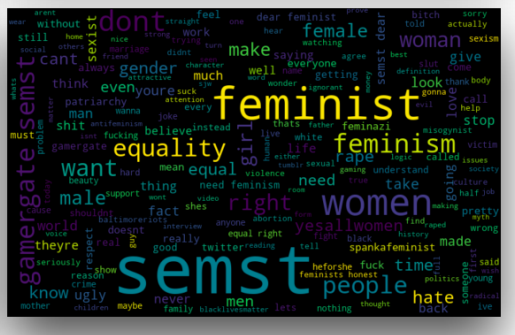
Legalization Of Abortion



Atheism



Climate Change

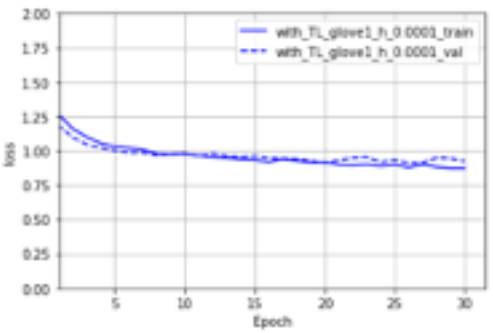


Feminist

#gohillaryHillary Clinton
#gohillary2016Hillary Clinton
#hillary2016Hillary Clinton
#whyiamnotvotingforhillaryHillary Clinton
#ohhillnoHillary Clinton
#stophillaryHillary Clinton
#hillaryHillary Clinton
#hillaryclintonHillary Clinton
#climatechangeClimate Change is a Real Concern
#globalwarmingClimate Change is a Real Concern
#climatechangescamClimate Change is a Real Concern
#globalwarminghoaxClimate Change is a Real Concern
#junkscienceClimate Change is a Real Concern
#globalcoolingClimate Change is a Real Concern
#globalwarmingisnotrealClimate Change is a Real Concern
#prochoiceLegalization of Abortion
#abortionLegalization of Abortion
#prolifeLegalization of Abortion
#praytoendabortionLegalization of Abortion
#EndAbortionLegalization of Abortion
#PlannedParenthoodLegalization of Abortion
#prayerAtheism
#faithAtheism
#religionAtheism
#atheismAtheism
#atheistAtheism
#JesusAtheism
#religionpoisoneverythingAtheism
#antireligionAtheism
#NoMoreReligionsAtheism
#AntiTheismAtheism
#antiatheistAtheism
#normalizeatheismAtheism
#ChristopherHitchensAtheism
#secularAtheism
#HumanismAtheism
#secularismAtheism
#FeminismFeminist Movement
#FeministsAreUglyFeminist Movement
#INeedFeminismBecauseFeminist Movement
#WomenAgainstFeminismFeminist Movement
#FeminismIsAwfulFeminist Movement

Fig 6

```
plotter(m_Target, ylim=[0.0, 2], metric = 'loss')   plotter(m_Target, ylim=[0.0, 1.1], metric = 'f1_m')
```




```
print(confusion_matrix(predict,Y_test))
```

```
[[516 137 139]
 [ 34  90  25]
 [165  77  66]]
```

Fig 8

lstm layer1 32,dropout 0.7

lstm layer2 32 dropout 0.7

kernal regulariser 0.0001

dropout level1 0.5

dropout level2 0.5

Fig 9

References

[1]	S. K. P. S. X. Z. C. C. Saif M. Mohammad1, "A Dataset for Detecting Stance in Tweets," [Online]. Available: https://www.aclweb.org/anthology/L16-1623.pdf. [Accessed 10 October 2020].
[2]	B. P. L. L. Matt Thomas, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," December 2006. [Online]. Available: https://www.cs.cornell.edu/home/llee/papers/tpl-convote.dec06.pdf. [Accessed 10 oct 2020].
[3]	S. M. M. e. al, "SemEval-2016 Task 6," June 2016. [Online]. Available: http://alt.qcri.org/semeval2016/task6/. [Accessed 11 Oct 2020].
[4]	S. K. S. Z. C. Saif M. Mohammad, "SemEval-2016 Task 6: Detecting Stance in Tweets," 2016. [Online]. Available: https://www.aclweb.org/anthology/S16-1003.pdf. [Accessed 11 Oct 2020].
[5]	R. S. S. K. Kuntal Dey, "Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention," 9 Jan 2018. [Online]. Available: https://arxiv.org/pdf/1801.03032.pdf. [Accessed 9 Oct 2020].
[6]	P. S. K. Saif M. Mohammad, "Stance and Sentiment in Tweets," 5 May 2016. [Online]. Available: https://arxiv.org/pdf/1605.01655.pdf. [Accessed 6 Oct 2020].
[7]	S. K. S. S. P. Z. X. C. Mohammad, "SemEval-2016 Task 6: Detecting Stance in Tweets," *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016),* p. 31–41, June 2016.
[8]	N. S. B. O. D. D. Kevin Gimpel, "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments," [Online]. Available: https://www.aclweb.org/anthology/P11-2008.pdf. [Accessed Oct 2020].
[9]	Y. Kim, "Convolutional Neural Networks for Sentence Classification," 3 september 2014. [Online]. Available: https://arxiv.org/pdf/1408.5882.pdf. [Accessed 9 october 2020].
[10]	X. Z. X. L. W. C. T. W. Wan Wei, "pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection," 2016. [Online]. Available: https://pdfs.semanticscholar.org/d16b/75ab2c4c5560212cea1eb9633abedfeaf4c5.pdf. [Accessed 8 oct 2020].
[11]	G. H. A. K. I. S. R. S. Nitish Srivastava, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," June 2014. [Online]. Available: https://jmlr.org/papers/volume15/srivastava14a.old/srivastava14a.pdf. [Accessed 8 Oct 2020].
[12]	G. Z. a. A. Marsh, "MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection," 13 June 2016. [Online]. Available: https://arxiv.org/pdf/1606.03784.pdf. [Accessed 10 October 2020].
[13]	N. K. P. Simona Frenda, "Stance or insults?," 10 june 2019. [Online]. Available: http://personales.upv.es/prosso/resources/FrendaEtAl_EVIA19.pdf. [Accessed 11 oct 2020].