Multi class twitter data classification

Falguni Vasava,Registration No:2006617, University of Essex, Department of Computer science and Electronics

Abstract—Recently many researchers have been attentive in studying online social media platforms as they are growing rapidly and the impact online social media content made on people's behaviour. most of the work focuses on the sentimental analysis and opinion mining. It is a field which automatic identifies the opinion of user towards a product or an organization or a person by analysing their posts and publications. The most popular microblogging service today that allows users to share their aspects of life and opinions is Twitter. Tweets are brief,(2021) 140-character messages that can be pull out to analyse user comments. (-Report), 2021)Tweets express the opinion and emotion of every individual. That being the case, opinion mining of twitter data has been given much attention over the past decade. As such, we present the model to classify the sentiment of various twitter data.

1 Introduction

TWITTER has turn out to be a common specialized communication tool between internet users, in which tens of thousands and thousands of status keep posted that voice critiques taking place loads of subjects or offer private state of mind, event and statements, are "tweeted" every solitary day. It additionally has turn out to be a critical platform for country wide discussions which lets in community, politicians and scholars to get a brand new and well-timed update of the general public critiques in the direction of an event like natural disaster or terror attack. (Mengdi Li, 2018)

Also, relationships between Twitter users are not necessarily reciprocal. Likewise, Twitter has a free-to-use API that allows anyone to tweet and work on any topic in real time.

That being the case, Scientists have more attention into text analysis of tweets (-Report), 2021). As it will be helpful in businesses to plan marketing strategies built on product and feedback. It can correspondingly be supportive to analyse public opinion on subject to politics and entertainment industry etc.

Text classification is the process of categorizing the text into an organized group by labelling them with relevant categories from a predefined set of categories. Scientists handled the text classification problem in many ways. Support Vector Machine (SVM)[1] ,Naïve Bayes (NB)[2], K Nearest Neighbours (KNN) [3], Probabilistic Bayesian models [4, 5], Decision Rules [6, 2], Decision Tree (DT) [7, 8], hidden Markov model (HMM), etc. are fundamentally commonly used text classification models. The aim of the proposed method is to categorized tweets into predefined classes.

2 BACKGROUND

CLASSIFICATION is a way of distributing documents to pre-established groups of classes and plays an imperative role in numerous data management applications. until now, several mainstream text categorization systems, for example Naïve Bayes, Neural Network, Decision Tree, k-Nearest Neighbour, Support Vector Machine, are effectively applied to text classification. (Pingpeng Yuan, 2008)

Berger and Ghani studied "Naïve Bayes" as the binary classifier for Error-correcting Output Codes for multi class problem on the dataset of 105 classes of industry sector, and gives the best decrease in error. (Pingpeng Yuan, 2008)

A software has been built up by "Pazzani" et al. that figured out how to grade pages are given on the www based on user ruling. They analysed 3 distinct calculations: "The Bayesian classifier", "Decision Tree" (DT) & "k-Nearest Neighbour" with a two fold feature vector on the 2 classes of user likings [9]. Their experimental outcomes demonstrated in which DT is not fit to their issue statement and the k-Nearest Neighbour classifier functioned admirably over different techniques when given countless samples.(Pingpeng Yuan, 2008)

"Ou" & "Murphey" [10] presented study in the multi-class neural learning, enforced analysing training time complexity related with different methodology by multi-class pattern classifier. (Pingpeng Yuan, 2008)

Dmitry Davidov, Oren Tsur, Ari Rappoport proposed a supervised of sentiment classification framework that relies on fact provided by twitter. They used feature vector and K-nearest neighbour. The fundamental reason for this system is to recognize and discriminate between sentiment types outlined by tags and smiley.[11]

(-Report), 2021) Hetu et al. anticipated one model in sentiment analysis on twitter data. They classify the individuals' sentiments grounded on positive side & negative side reviews. This model gives high precision arranged huge dataset [12].

3 METHODOLOGY

THE general architecture of text cataloguing is shown in the figure. This design can be utilized as a fundamental /starting model for many classification tasks.

(-Report), 2021) Cleaning up your data set is an important step in machine learning. Numerous Toy Datasets do not have to clean, considering the fact is that it's up to now spotless, peer-assessed & distributed in a way you can employed it is precisely to deal with the learning calculations. But for most of the data there are some unhelpful information present. This can affect the final result of the classifier. The basic steps in data cleansing are symbol removal, URL removal, number removal, and more. After

Fig. 1: Architecture of text classification

that, the features are pull out from the data. Can be used to generate a classifier model. Next step is to detached training dataset and test dataset. So that test dataset can additionally be used for classifier assessment or class prediction of new text. By feeding features of training data and their label as an input classifier is trained in the next step. Trained classifier further be used to evaluate the classifier prototype or predict label of text presented in assessment data.

3.1 Data cleaning

Twitter data might hold bunch of noise and uninformative parts like digits, URL's, handles, symbols and so on. By keeping these noise and uninformative fragments of the text may get the issue with high dimensional. Therefore, the classification becomes extra troublesome since every word in the text is pickled as one dimension. Furthermore, that unproductive parts might influence on accurateness of model.

Proposed text cleaning process encompasses several steps: 1) coverts all uppercase character to lowercase,2) replace symbol with the white space,3) remove punctuation and numbers,4) remove single character, 5) remove multiple space, 6) remove stop words.

Separating words from a text string is called tokenization. And stemming is to removing the suffix of the word.

3.2 Feature Engineering

(-Report), 2021) The data cleaned up at this stage (pre-processed data) is transmuted into flat features that can further be applicable in machine learning model. Proposed method use tokenizer function from keras pre-processing library.

3.3 Train classifier

The extracted features can be used to train the classifier. Several machine learning models would be used for text classification. For illustration: baggie models, naïve bayes classifier, support vector machine, linear classifier, etc.

Proposed model uses keras neural network sequential model for classification.

4 Results

4.1 Dataset

The proposed model uses TweetEval [13] dataset for multi class text classification of data provided by twitter.. (-Report), 2021) There are 7 classes in TweetEval dataset. Proposed method use three classes(irony recognition, Hate speech detection, Offensive language identification) from them.(-Report), 2021) Table 2 contains the TWEETEVAL datasets statistics.

TABLE 1: Dataset

Task	Lab	Train	Val	Test
Irony detection (-Report), 2021)	2	2862	955	784
Hate speech det. (-Report), 2021)	2	9000	1000	2970
Offensive lg. id.	2	11916	1324	860

Sample text data are shown in Table.2 with their true label.

sample data are shown in table 3 after applying text cleaning function.

Summary of the model created by using keras neural network is shown in figure below. (for irony)

```
Model: "sequential_11"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_21 (Dense)             (None, 512)               512512
_________________________________________________________________
activation_20 (Activation)   (None, 512)               0
_________________________________________________________________
dropout_11 (Dropout)         (None, 512)               0
_________________________________________________________________
dense_22 (Dense)             (None, 2)                 1026
_________________________________________________________________
activation_21 (Activation)   (None, 2)                 0
=================================================================
Total params: 513,538
Trainable params: 513,538
Non-trainable params: 0
_________________________________________________________________
None
```

Training Accuracy and loss graphs for each task are shown below:

TABLE 2: Raw sample data text

No	Data	Lable
1	seeing ppl walking w/ crutches makes me really excited for the next 3 weeks of my life (-Report), 2021)	irony
2	look for the girl with the broken smile, ask her if she wants to stay while, and she will be loved. (-Report), 2021)	Not-irony
3	Now I remember why I buy books online @user #ser-vicewithasmile	irony
4	@user @user So is he banded from wearing the clothes? #Karma	irony
5	Just found out there are Etch A Sketch apps. #old-school #notoldschool	irony

TABLE 3: cleaned sample data text

No	Data
1	nice new signage. Are con-cerned Beatlemania -style hys-terical crowds crongregating
2	nice new signage. Are con-cerned Beatlemania -style hys-terical crowds crongregating
3	3real talk eyes gouged rapefugee?
4	girlfriend lookin like groupie bitch!
5	Hysterical woman like

HATE:

Fig. 2: Accuracy graph of hate task

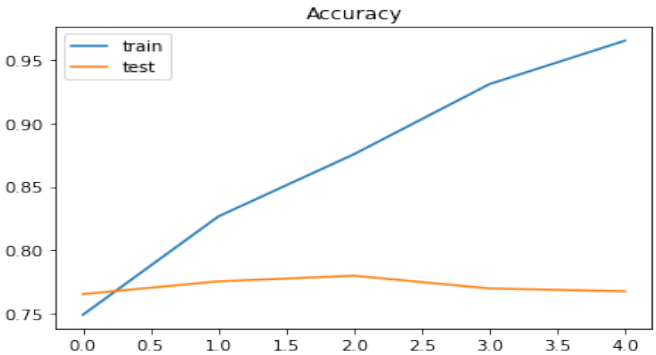
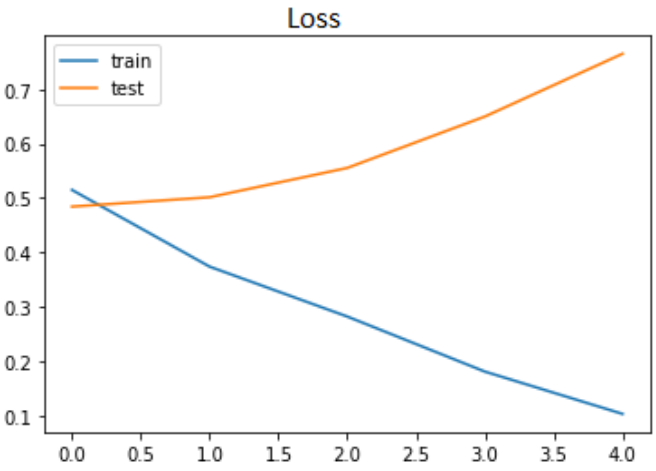


Fig. 3: loss graph of hate task



IRONY:

Fig. 4: Accuracy graph of irony task

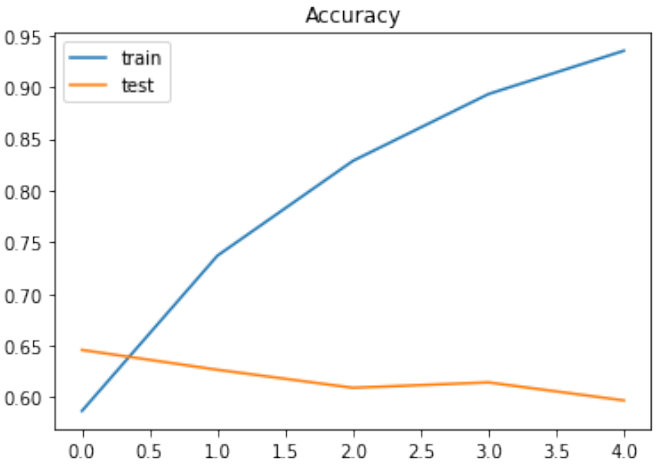


Fig. 5: loss graph of irony task

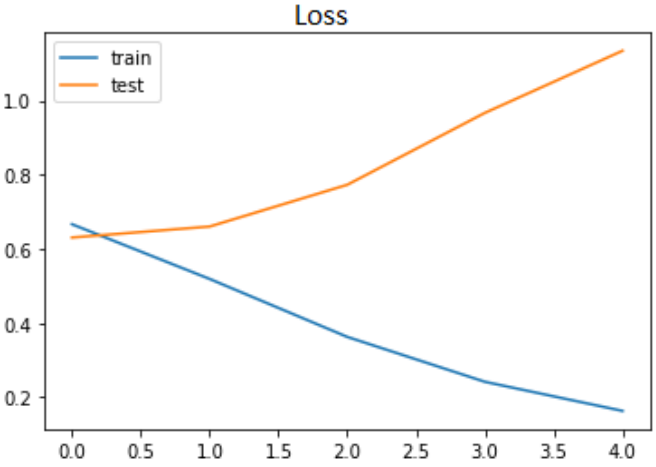


Fig. 6: Accuracy graph of Offensive task

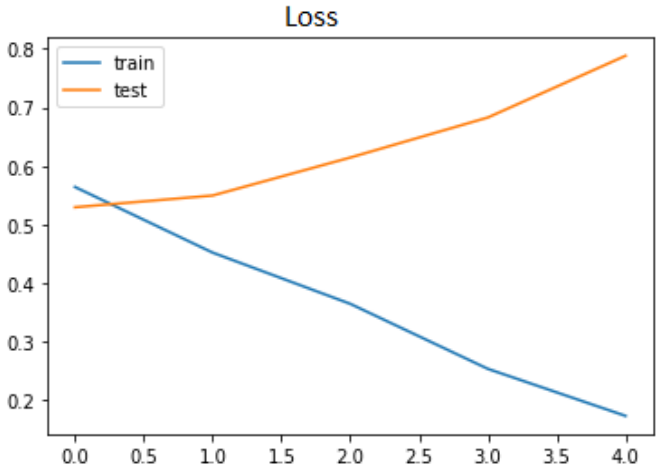
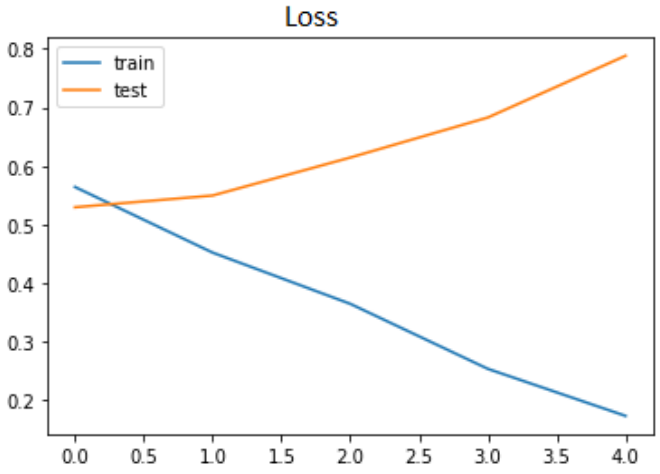


Fig. 7: loss graph of Offensive task



OFFENSIVE:

5 DISCUSSION

An assessment matric computes the performance of a predictive model. For classification problems, metrics compares the anticipated class label to the prophesied class label. For testing the performance of this study, we will be using F1-score. It wholes precision and recall procedures under the concept of harmonic mean.

$$F1 - Score = \left(\frac{2}{precision^{-}1 + recall^{-}1}\right) \tag{1}$$

$$F1 - Score = 2\left(\frac{precision * recall}{precision + recall}\right) \tag{2}$$

The formulation of F1-score can be considered as a weighted average between precision and recall, where F1-score reaches its best value at 1 and worst score at 0.

When we use F1 score to measure performance of multi-class classifier, F1-score should involve all the classes.

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{3}$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{4}$$

Above formula represents the two quantities(precision and recall) for a generic class k.

After that, macro average precision and macro average recall are simply computed as the arithmetic mean of the metrics for single classes.

$$MacroAveragePrecision = \frac{\sum_{k=1}^{K} Precision_k}{K} \tag{5}$$

$$MacroAverageRecall = \frac{\sum_{k=1}^{K} Recall_k}{K} \tag{6}$$

Lastly, macro F1-score is the harmonic mean of macro average precision and macro average:

$$MacroF1 - Score = 2 * \left(\frac{MAP * MAR}{MAP + MAR}\right) \tag{7}$$

The evaluation of created model is done by TweetEval comparative evaluation[20] The result of each task is presented in table 5.

TABLE 4: F1-score and accuracy of each task

No	Task	Macro-f1	Accuracy
1	Hate	38.92	52.55
2	Offensive	69.66	77.67
3	Irony	63.39	63.26

Accuracy comparison of proposed algorithm with other models is presented in table below.

TABLE 5: model comparison

Model	Irony	Hate	Offensive
RoBERTa-Retrained	61.7	52.3	80.5
RoBERTa-Base	59.7	46.6	79.5
RoBERTa-Twitter	65.4	49.9	77.1
FastText	63.1	50.6	73.4
LSTM	62.8	52.6	71.7
SVM	61.7	36.7	52.3
Proposed	63.39	52.55	77.69

6 CONCLUSION

TWITTER has become a significant site for public conversations where people in general, researchers, and government officials can get a novel and timely update of the public opinion in the direction of an occurrence like terror attack or election. To automatically excavation and perceive the sentiment and opinions people are communicating, Multi-class text classification can be useful method.

This report presents the keras neural network model for twitter data that include three tasks(hate, irony, offensive). (Mengdi Li, 2018)

REFERENCES

[1] Bing Liu, Wynne Hsu, and Yiming Ma. 1998. Integrating classification and association rule mining. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98). AAAI Press, 80–86.

[2] Lewis, D. D. (1998), Naive (Bayes) at forty: The independence assumption in information retrieval., in Claire Nédellec Céline Rouveirol, ed., 'Proceedings of ECML-98, 10th European Conference on Machine Learning' , Springer Verlag, Heidelberg, DE, Chemnitz, DE , pp. 4–15 .

[3] Osmar R. Zaïane and Maria-Luiza Antonie. 2002. Classifying text documents by associating terms with text categories. In Proceedings of the 13th Australasian database conference - Volume 5 (ADC '02). Australian Computer Society, Inc., AUS, 215–222.

[4] Hang Li and Kenji Yamanishi. 1999. Text classification using ESC-based stochastic decision lists. In Proceedings of the eighth international conference on Information and knowledge management (CIKM '99). Association for Computing Machinery, New York, NY, USA, 122–130. DOI:https://doi.org/10.1145/319950.319966

[5]] Wenmin Li, Jiawei Han and Jian Pei, "CMAR: accurate and efficient classification based on multiple class-association rules," Proceedings 2001 IEEE International Conference on Data Mining, 2001, pp. 369-376, doi: 10.1109/ICDM.2001.989541.

[6] Maria-Luiza Antonie, Osmar R. Zaïane, and Alexandru Coman. 2001. Application of data mining techniques for medical image classification. In Proceedings of the Second International Conference on Multimedia Data Mining (MDMKDD'01). Springer-Verlag, Berlin, Heidelberg, 94–101.

[7] Hu, Zhi-kun Gui, Wei-hua Yang, Chunhua Deng, Peng-cheng Ding, Steven. (2011). Fault Classification Method for Inverter Based on Hybrid Support Vector Machines and Wavelet Analysis. International Journal of Control, Automation and Systems. 9. 797-804. 10.1007/s12555-011-0423-9.

[8] Joachims T. (1998) Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg.

[9] M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill Webert: identifying interesting web sites", Proc. of the Thirteenth Amer. Nat. Conf. on Artificial Intelligence (AAAI-96), Vol.1. AAAI Press, Portland, OR, 1996, pp.54-61

[10] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks", Pattern Recognition, Vol.40, No.1, January 2007, pp.4-18

[11] Davidov D., Tsur O., Rappoport A." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys".

[12] Hetu Bhavsar, Richa Manglani" Sentiment Analysis of Twitter Data using Python"International Research Journal of Engineering and Technology (IRJET) Mar 2019e-ISSN: 2395-0056 p-ISSN: 2395-0072

[13] https://github.com/cardiffnlp/tweeteval