Evaluate the Tweet Analysis with Improved Accuracy Using Multi Channel N-gram Convolutional Neural Network Model over Naive Bayes Model

Chinthapalli Satya Swaroop Reddy and P. Sriramya
Department of Computer Science Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, 602105, India

Keywords: Deep Learning, Neural Networks, Embedding, Tweets, Disaster Management, Naive Bayes, Novel Multi Channel N-gram CNN Model, Naive Bayes Model.

Abstract: The purpose of this study is to compare the accuracy of tweet analysis using a novel Multi Channel N-gram CNN model and Naive Bayes model. Materials and Methods: There are two groups in this study: Naive Bayes methods and Multi channel N-gram CNN. The sample size for each group is 10, and the study's parameters include an alpha value of 0.8 and a beta value of 0.2. Taking the G-Power value of 80% into account, the significance value of the dataset was predicted using SPSS. Results and Discussion: In the examination of tweets, the Multi Channel N gram CNN Algorithm's accuracy was 97.84%, whereas the Naive Bayes algorithm's accuracy was 79.69%; this means that the two algorithms are statistically different. Conclusion: When analyzing tweets, the Multi Channel N gram CNN algorithm performs noticeably better than the Naive Bayes algorithm.

1 INTRODUCTION

Over the past few generations with the increase in the usage of internet and mobile phones the usage of social media has increased to a high level. Any kind of general information can be shared in social media and that information can be viewed by people all around the world. Social media apps like Twitter, Instagram, Facebook and snapchat became very popular in the world. People can share their feelings and opinions on social media on different issues (Ninan, 2022).Any kind of disaster can occur at any place at any time. In the time of disaster, social media is the powerful platform that can spread the news about disaster all over the world within very less time. The information shared on social media may be useful for social welfare organisations, Disaster Management Teams, self help groups and rescue organisations as it can alert them to take the safety measures in advance(“Dormant Disaster Organizing and the Role of Social Media”, 2019). As there is no limit in sharing information in social media, there is a chance that many unwanted tweets will also be shared. People will be sharing both real news and fake news on social media. So proper analysis of tweets is required and it plays an important role in many activities (“Multimodal Analysis of Disaster Tweets” n.d., 2021). Social media is capable of spreading information within a fraction of seconds all over the world, so many people agree that social media is a powerful tool that can make people and Disaster Management Teams aware of all situations that are happening around the world. (Maulana and Maharani, 2021) (Deena, S. et al., 2022). Different comments and tweets can have a negative impact on the information itself, which makes it difficult for many rescue and emergency responders to develop efficient knowledge management strategies for a catastrophe setting that is rapidly shifting (Hadiana and Ningsih, 2021).

Many organisations, Disaster Management Teams and people found the importance of proper analysis of tweets and they started working on generating new models that can analyse the tweets and can give accurate outcomes. With the increase in work on analysis of tweets, many articles are also published on disaster tweet analysis across various journal databases likeScienceDirect, IEEE, E Village, etc. Total of 527 articles were published in ScienceDirect’s database on Disaster tweet analysis

Reddy, C. and Sriramaya, P.
Evaluate the Tweet Analysis with Improved Accuracy Using Multi Channel N-gram Convolutional Neural Network Model over Naive Bayes Model.
DOI: 10.5220/0012772600003739
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 546-552
ISBN: 978-989-758-661-3
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

in IEEE database using machine learning algorithms over the last 5 years and around and 13 Journals were published in IEEE database. Hien et al compared learning-based and matching-based techniques for finding related tweets. and came to the conclusion that, as compared to the learning-based approach, the matching-based methodology includes tweets that are higher in quality but less relevant 2017 (To et al.). J. Rexiline Ragini collected the data from Twitter about disaster tweets and the analysis of big data is done using Apache Spark Big Data Framework and Python programming language(Sitaula and Shahi, 2022) (Ramkumar. G et al., 2022). Shamanth Kumar from Arizona State University introduced a Tweet Tracker application that constantly monitors Twitter streaming feed using particular hashtags and keywords about disaster (Kumar et al., 2011). An innovative method to visualize the emotional state of the general public in the event of a natural disaster was given by (Shekhar and Setty, 2015).

There are a lot of research articles published on proper analysis of tweets. In all the papers the authors used different algorithms and models in both Machine Learning and in Deep Learning, but many of the results in predicting accuracy were lower than expected so it became the major drawback in almost every research paper (Maulana and Maharani, 2021).

So, this research study aims to improve the accuracy of predicting whether the tweet is real or fake with the help of Multi Channel N gram CNN algorithms in Machine Learning with less computational time.

2 MATERIALS AND METHODS

Saveetha School Of Engineering in Saveetha University provided its Data Science lab for this research work for study setting. The main Objective of this research is to do a comparative analysis on two groups. First group is Multi Channel N-gram CNN algorithm and second group is Naive Bayes algorithm. Same sample size of 10 is taken for each of the two groups (“Tweet Analysis - ANN/BERT/CNN/n-Gram CNN”, 2020). For the purposes of this study, experimental computation is carried out utilizing G-power as 80% with a 95% confidence interval, alpha as 0.05, and beta as 0.2. The data set Sample_Submission.csv for this comparative study is taken from the open platform Kaggle.com.

2.1 Multi Channel N-gram CNN Model

A multi-channel CNN model is a convolutional neural network that takes input from multiple channels or sources. Each channel represents a different aspect or feature of the input data.

Multiple convolutional layers are often followed by pooling layers and fully linked layers in the design of a multi-channel CNN model. Each channel is fed into a separate set of convolutional layers, and the outputs from each set of layers are combined before being passed to the next set of layers. The advantage of using a multi-channel CNN model is that it allows the system to learn different aspects of the input data separately and then combine them to make a more accurate prediction. It can also help to reduce overfitting by providing multiple sources of information to the network. Overall, a multi-channel CNN model is a powerful tool for tasks that involve complex input data with multiple sources of information. This Multi Channel CNN approach was first used by Yoon Kim in his paper titled “Convolutional Neural Networks for Sentence Classification” (Kim, 2014).

To perform tweet analysis prediction using Multi Channel CNN, the following steps are involved:

1. Encrypt the data
2. Define Model
3. Fitting data in the Model
4. Predict the outcome of text data.

2.2 Naive Bayes Model

The Naive Bayes model is used mostly for solving classification problems using a probabilistic approach. This model is based on the popular mathematical theory called the Bayes probability theorem. In the case of Bayes theorem, the occurrence of one event is always independent of occurring other events and so it is called naive. The Naive Bayes algorithm is expected to show optimal prediction having a high range of applicability compared with other models. This classifier has various applications as it is used in many problems like classification problems, sentimental analysis, fraud detection e.t.c. (Ji, Yu, and Zhang, 2011). The formula for Bayes theorem is stated as below:

$$P(A|B) = P(B|A)*P(A)/P(B)$$

Where ,

- P(A|B) is Probability of occurrence of event A after B
- P(B|A) is c after A
- P(A) is Probability of occurrence of event A
- P(B) is Probability of occurrence of event A

The dataset sample_submission.csv is used in this study. The sample_submission was divided into two different parts in the proportion of 80/20. Major part is used for training purposes and the minor part is used for the testing process. Names of both datasets are train.csv and test.csv respectively. By using both the training and testing datasets the algorithm was implemented to get the outcome. The laptop with an Intel i5 processor, 8GB of RAM, 64-bit Windows 11 operating system, and other features is used to conduct this research.

2.3 Statistical Analysis

The software utilized in this instance is IBM SPSS V22.0 for statistical implementation. For statistical computations like mean and standard deviation as well as to help layout the graphs, we use the Statistical Package for Social Sciences (SPSS). The TweetsNumber and DataSize are the independent variables. 'Accuracy' is the dependent variable. For each group, a sample size of 10 is used to construct the dataset, and accuracy is used as the testing variable. An independent samples t-test is performed to statistically compare the two methods' significance.

3 RESULTS

The ultimate goal of this research article is to compare the accuracy in analysis of tweets using the Multi Channel N-gram CNN model and Naive Bayes. The most accurate algorithm is selected between the two algorithms based on the accuracy of its output. The accuracy shown by the Multi Channel N-gram CNN model is 97.84% whereas the accuracy shown by the Naive Bayes model is 79.69%.

Table 1 represents the sample data set taken for this research work.

Table 2 represents the Pseudocode for Multi Channel N-gram CNN model.

Table 3 represents the Pseudocode for the Naive Bayes model.

Table 1: SampleDataset.

id	text	target
1	Our actions are what caused this earthquake, I want Allah to pardon us all.	1
4	Canadian forest fire near La Ronge, Saskatchewan	1
5	Officers have requested that all residents "shelter in place." There aren't any further anticipated evacuation or stay-in-place orders.	1
6	13,000 residents in California are issued evacuation orders due to wildfires.	1
7	Just received this picture from Ruby, Alaska, showing smoke from wildfires entering a school.	1
8	California Highway 20 is closed in both directions due to a fire in Lake County (#RockyFire Update) - #CAfire #wildfires	1
10	#disaster #flood Flash flooding is caused by heavy rain in the Manitou and Colorado Springs areas.	1
13	The fire in the woods is visible from where I am standing on the hilltop.	1
14	Since the building across the street is currently undergoing an emergency evacuation,	1
15	I'm worried that a tornado will soon hit our neighborhood.	1

Table 2: Pseudocode for Multi Channel N-gram CNN model.

// I: Input dataset records
1. Import the required packages.
2. Convert the string values in the dataset to numerical values.
3. Assign the data to X_train, y_train, X_test and y_test variables.
4. Using train_test_split() function, pass the training and testing variables and give test_size and the random_state as parameters.
5. Import the Multi Channel N-gram CNN model.
6. Using the Multi Channel N-gram CNN model, predict the output of the testing data.
7. Calculate the accuracy
OUTPUT **//Accuracy**

Table 3: Pseudocode for Naive Bayes model.

```
// I: Input dataset records
1. Import the required packages.
2. Convert the string values in the dataset to numerical values.
3. Assign the data to X_train, y_train, X_test and y_test variables.
4. Using train_test_split() function, pass the training and testing variables and give test_size and the random_state as parameters.
5. Import the Naive Bayes model.
6. Using the Naive Bayes model, predict the output of the testing data.
7. Calculate the accuracy
OUTPUT
//Accuracy
```

Table 4: Accuracy of Classification of Tweet analysis using Multi Channel N-gram CNN model.

GROUP	ACCURACY	LOSS
TEST 1	97.83	2.17
TEST 2	97.76	2.24
TEST 3	97.93	2.07
TEST 4	97.88	2.12
TEST 5	97.76	2.24
TEST 6	97.88	2.12
TEST 7	97.79	2.21
TEST 8	97.81	2.19
TEST 9	97.93	2.07
TEST 10	97.84	2.16

Table 5: Accuracy of Classification of Tweet analysis using Naive Bayes model.

GROUP	ACCURACY	LOSS
TEST 1	79.62	20.38
TEST 2	78.72	21.28
TEST 3	79.56	20.44
TEST 4	79.67	20.33
TEST 5	79.14	20.86
TEST 6	80.30	19.70
TEST 7	80.25	19.75
TEST 8	80.40	19.60
TEST 9	79.25	20.75
TEST 10	80.04	19.96

Table 6: Group Statistics Results represented for Accuracy for Multi Channel N-gram CNN and Naive Bayes algorithms.

Algorithm	N	Mean	Std. Deviation	Std. Error Mean
Accuracy N-gram CNN	10	97.8410	0.06297	0.01991
Naive Bayes	10	79.6950	0.55490	0.17548

Table 7: Independent Samples T-test shows significance value achieved is p=0.000 (p<0.05), which shows that the two groups are statistically significant.

	Levene's test for equality of variances		T test for Equality of means						
	F	Sig	t	df	Sig(2-tailed)	Mean Difference	Std Error Difference	95% confidence level Lower	95% confidence level Upper
Accuracy Equal variances assumed	16.619	0.01	102.751	18	0.000	18.146	0.17660	17.77497	18.51703
Accuracy Equal variances not assumed			102.751	9.232	0.000	18.146	0.17660	17.74802	18.54398

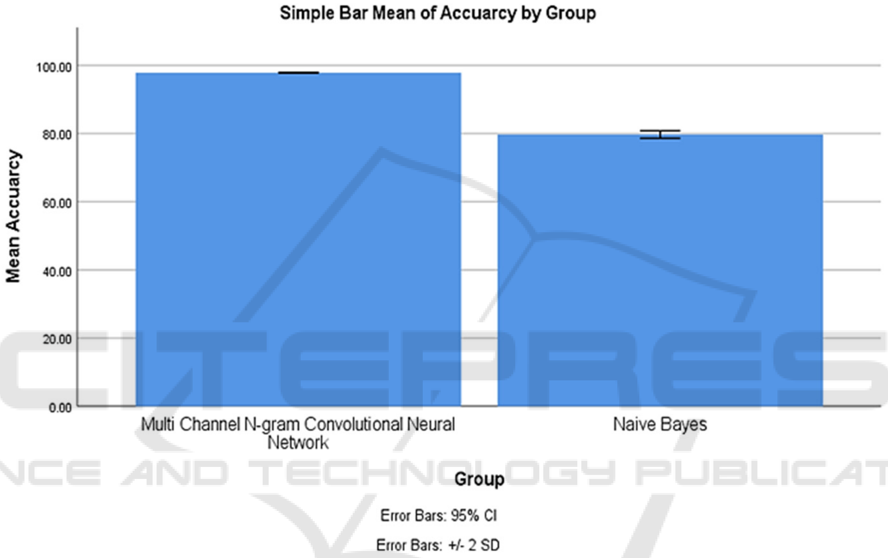

Figure 1: Bar chart showing the comparison of Multi Channel N gram CNN (97.84%) and Naive Bayes (79.69%) in terms of mean accuracy. X-Axis: Multi Channel N gram CNN (N gram CNN) VS Naive Bayes and Y-Axis: the Mean accuracy of detection with ±2 SD.

Table 6 shows the Group statistics results represented for Accuracy and Loss for Multi Channel N-gram CNN model and Naive Bayes model. The mean, standard deviation and standard error mean for Multi Channel N-gram CNN model is 97.84,0.06297 and 0.01991 respectively. The mean, standard deviation and standard error mean for Naive Bayes model is 79.6950,0.55490 and 0.17548 respectively. So by comparing the results it is very much clear that the Multi Channel N-gram CNN model is more accurate than Naive Bayes model in analysing the tweets.

Table 7 shows the independent sample T test performed on Multi Channel N-gram CNN model and Naive Bayes model to calculate the accuracy and loss in both equal variance assumed and equal variance not assumed. With the confidence level of 95% it also gave the values of mean difference and standard error difference.

Figure 1 shows the comparison of both algorithms with the help of a bar graph. The bar graph is plotted in between Multi Channel N-gram CNN model and Naive Bayes model. Accuracy is taken on the X axis and Group names are taken on the Y axis in the bar graph. By observing the bar graph we can understand that there is a significant difference in between both the algorithms in terms of accuracy. The accuracy of the Multi Channel N-gram CNN model is more when compared to Naive Bayes model.Total of 10 iterations were performed on both the proposed model and existing model and all the outcomes are noted in Table 4 and Table 5. An Independent Sample Test was performed using the SPSS tool.

4 DISCUSSIONS

By comparing all the outcomes and results it is observed that the Multi Channel N-gram CNN model is showing much more accurate results in analysis of disaster tweets than Multi Channel N-gram CNN model. The accuracy of the Multi Channel N-gram CNN model is 97.84%, loss is 2.16% and the accuracy of Glove with Keras Word embedding model is 55.06%, loss is 46.94%.

Some of the research articles that are already published are supportive to our research article. Author proposed a model using Multi Channel CNN model for classifying the covid related tweets and got the accuracy of 94.56%.(Sitaula and Shahi 2022).Author proposed a model for analysing the disaster related images using the Multi Model network, VCG-16, ResNet-50 and Xception Network. And he concluded that analysis of disaster related images is best done using the Multi Model network (Asif et al. 2021). Author proposed a model to analyse the disaster tweets using CNN and ANN algorithms, and he also concluded that accuracy was better in both CNN and ANN combined than the individual algorithms.(Mathur, Sharma, and Veer 2022). Author in his research work used Naive Bayes algorithm, CNN with Multi channel distribution and CNN without Multi channel distribution for classifying the disaster tweets. And he concluded that analysis of tweets using CNN with a multi-channel model gave highly accurate results.(Sitaula and Shahi 2022). Limitations of our work is that this method is feasible on the offline datasets of significant size and the live updates can not be known using this analysis. So the study was restricted to the limit of data availability that might contain only some part of disaster related tweets. The prediction done by the algorithm may be much more different than the real time live prediction. Future scope of this study is I intend to extend our database to the other networking apps like facebook, instagram etc. I also intend to add disaster prediction models to work to know the trends of disasters in various regions. By further developing the work, it might be very useful to various disaster management teams and organisations.

5 CONCLUSION

In this research work, the results show us that Multi Channel N-gram CNN model can be used in the analysis of disaster tweets with improved accuracy of 97.84% than the Glove with Keras Word embedding model with accuracy of 55.06%.

REFERENCES

Asif, Amna, Shaheen Khatoon, 2021. Md Maruf Hasan, Majed A. Alshamari, Sherif Abdou, Khaled Mostafa Elsayed, and Mohsen Rashwan. “Automatic Analysis of Social Media Images to Identify Disaster Type and Infer Appropriate Emergency Response.” *Journal of Big Data* 8 (1): 1–28.

Deena, S. R., Kumar, G., Vickram, A. S. Singhania, R. R., Dong, C. D., Rohini, K., ... & Ponnusamy, V. K., 2022. Efficiency of various biofilm carriers and microbial interactions with substrate in moving bed-biofilm reactor for environmental wastewater treatment. Bioresource technology, 359, 127421.

“Dormant Disaster Organizing and the Role of Social Media.”, 2019. *New Media in Times of Crisis*. https://doi.org/10.4324/9780203703632-14.

Ji, Yaguang, Songnian Yu, and Yafeng Zhang, 2011. “A Novel Naive Bayes Model: Packaged Hidden Naive Bayes.” *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*. https://doi.org/10.1109/itaic.2011.6030379.

Kim, Yoon, 2014. “Convolutional Neural Networks for Sentence Classification,” August. https://doi.org/10.48550/arXiv.1408.5882.

Kumar, 2011, Shamanth, Geoffrey Barbier, Mohammad Abbasi, and Huan Liu. “TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief.” *Proceedings of the International AAAI Conference on Web and Social Media* 5 (1): 661–62.

A. Kishore Kumar, 2022. M. Aeri, A. Grover, J. Agarwal, P. Kumar, and T. Raghu, “Secured supply chain management system for fisheries through IoT,” *Meas. Sensors*, vol. 25, no. August, p. 100632, 2023, doi: 10.1016/j.measen.2022.100632.

Mathur, Prerak, Tanu Sharma, and Karan Veer, 2022. “Analysis of CNN and Feed Forward ANN Model for the Evaluation of ECG Signal.” *Current Signal Transduction Therapy*. https://doi.org/10.2174/1574362417666220328144453

Maulana, Iqbal, and Warih Maharani, (2021. “Disaster Tweet Classification Based On Geospatial Data Using the BERT-MLP Method.” *2021 9th International Conference on Information and Communication Technology (ICoICT)*. https://doi.org/10.1109/icoict52021.2021.9527513.

“Multimodal Analysis of Disaster Tweets.” n.d. Accessed December 19, 2022. https://ieeexplore.ieee.org/document/8919468.

Ninan, Johan., 2022. “The Past, Present and Future of Social Media in Project Management.” *Social Media for Project Management*. https://doi.org/10.1201/9781003215080-1.

Ningsih, A. K., and A. I. Hadiana, 2021. “Disaster Tweets Classification in Disaster Response Using Bidirectional

Encoder Representations from Transformer (BERT).” *IOP Conference Series: Materials Science and Engineering* 1115 (1): 012032.

Ramkumar, G. et al., 2021. “An Unconventional Approach for Analyzing the Mechanical Properties of Natural Fiber Composite Using Convolutional Neural Network” Advances in Materials Science and Engineering vol. 2021, Article ID 5450935, 15 pages, 2021. https://doi.org/10.1155/2021/5450935

Shekhar, Himanshu, and Shankar Setty, 2015. “Disaster Analysis through Tweets.” In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. https://doi.org/10.1109/icacci.2015.7275861.

Sitaula, Chiranjibi, and Tej Bahadur Shahi, 2022. “Multi-Channel CNN to Classify Nepali Covid-19 Related Tweets Using Hybrid Features,” March. https://doi.org/10.48550/arXiv.2203.10286.

To, Hien, 2017, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. 2017. “On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based?” In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*. IEEE. https://doi.org/10.1109/bigmm.2017.82.

“Tweet Analysis - ANN/BERT/CNN/n-Gram CNN”, 2020. Kaggle. July 19, 2020. https://kaggle.com/code/jagdmir/tweet-analysis-ann-bert-cnn-n-gram-cnn.

V. P. Parandhaman, 2023. "An Automated Efficient and Robust Scheme in Payment Protocol Using the Internet of Things," Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-5, doi: 10.1109/ICONSTEM56934.2023.10142797.