Evaluation of Multi-Channel N-gram Convolutional Neural Network for Improved Tweet Analysis Accuracy

Chinthapalli Satya Swaroop Reddy and P. Sriramya
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, India

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

Abstract: This study aimed to juxtapose the efficacy of the innovative Multi-Channel N-gram CNN model with the Naive Bayes model in tweet analysis. Two groups were established: the Naive Bayes and the Multi-Channel N-gram CNN, with each having a sample size of 10. The research parameters set were an alpha value of 0.8 and a beta value of 0.2. With a G-Power value of 80%, the significance of the dataset was ascertained using SPSS. Our findings highlighted that the Multi-Channel N-gram CNN algorithm achieved an accuracy of 97.84%, markedly outperforming the Naive Bayes which managed 79.69%. Consequently, for tweet analysis, the Multi-Channel N-gram CNN model is evidently superior.

1 INTRODUCTION

Over the past few generations, due to the rise in internet and mobile phone use, the popularity of social media has surged significantly. Social media platforms such as Twitter, Instagram, Facebook, and Snapchat have become global sensations. People frequently share their feelings and views on various topics on these platforms (Ninan 2022). Disasters can strike anywhere, anytime, and during such events, social media proves a potent tool for disseminating information swiftly across the globe. Such information can be invaluable to social welfare organisations, disaster management teams, self-help groups, and rescue organisations, providing early alerts for preventive action ("Dormant Disaster Organising and the Role of Social Media" 2019). However, the boundless nature of content sharing on social media also poses a risk of circulating misinformation. It becomes crucial, therefore, to rigorously analyse tweets, a task integral to numerous operations ("Multimodal Analysis of Disaster Tweets" n.d., 2021). Given its rapid dissemination capability, many acknowledge the power of social media in keeping people and disaster response units informed about global events (Maulana and Maharani 2021; Deena, S et al. 2022). But various tweets and comments can sometimes distort the very information they convey, posing challenges for rescue and emergency personnel trying to formulate effective strategies in dynamic disaster scenarios (Hadiana and Ningsih, 2021).

Several organisations and individuals recognise the imperative of tweet analysis and are developing innovative models for precise outcomes. A surge in tweet analysis research is evidenced by numerous publications across journal databases such as ScienceDirect, IEEE, and E Village. In the past five years, 527 articles were published on disaster tweet analysis using machine learning algorithms in the ScienceDirect database, with 13 journals appearing in the IEEE database. Hien et al. compared learning-based and matching-based methods for tweet relevance and deduced that while the matching-based approach yielded higher-quality tweets, they were less relevant (To et al. 2017). J. Rexiline Ragini utilised the Apache Spark Big Data Framework and Python for analysing disaster tweets sourced from Twitter (Sitaula and Shahi 2022; Ramkumar. G et al. 2022). Shamanth Kumar from Arizona State University devised a Tweet Tracker application monitoring Twitter's streaming feed using specific hashtags and keywords related to disasters (Kumar et al. 2011). Shekhar and Setty (2015) introduced a novel means to visualise public sentiment during natural calamities.

Despite the wealth of research on tweet analysis, many studies, employing both Machine Learning and Deep Learning techniques, have reported lower than anticipated accuracy results, a recurring limitation in most (Maulana and Maharani 2021). Thus, this research endeavour seeks to enhance the accuracy of discerning genuine tweets from spurious ones,

Reddy, C. and Sriramya, P.
Evaluation of Multi-Channel N-gram Convolutional Neural Network for Improved Tweet Analysis Accuracy.
DOI: 10.5220/0012544300003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 51-57
ISBN: 978-989-758-661-3

employing the Multi-Channel N-gram CNN algorithm in Machine Learning, while also aiming for reduced computational time.

2 MATERIALS AND METHODS

The Data Science lab at Saveetha School of Engineering, part of Saveetha University, graciously facilitated this research work by providing its facilities. This study's primary objective is a comparative analysis between two distinct groups: the Multi-Channel N-gram CNN algorithm and the Naive Bayes algorithm. Both groups were assigned an identical sample size of 10 (“Tweet Analysis - ANN/BERT/CNN/n-Gram CNN” 2020). The experimental computations for this investigation employed a G-power of 80%, a confidence interval of 95%, an alpha value of 0.05, and a beta of 0.2. The dataset, titled 'Sample_Submission.csv', utilised for this comparative research, was sourced from the publicly accessible platform, Kaggle.com.

2.1 Multi-Channel N-gram CNN Model

The multi-channel CNN model is a type of convolutional neural network that processes input from multiple channels or sources. Each channel encapsulates a unique facet or feature of the data. In constructing a multi-channel CNN model, convolutional layers typically precede pooling layers and fully connected layers. Every channel is channelled into its distinct convolutional layers, and the outcomes from these layers are amalgamated before progressing to subsequent layers. The merit of employing a multi-channel CNN model lies in its capacity to independently learn disparate aspects of input data and later consolidate them for a more precise prediction. Furthermore, it may curtail overfitting by presenting the network with a diversity of data sources. Altogether, a multi-channel CNN model proves invaluable for assignments entailing intricate input data with various informational facets. Yoon Kim was the pioneer in employing this Multi-Channel CNN approach, as delineated in his paper "Convolutional Neural Networks for Sentence Classification" (Kim 2014).

To undertake tweet analysis prediction via Multi-Channel CNN, the subsequent steps are essential:

1. Encrypt the data.
2. Outline the model.
3. Accommodate the data within the model.
4. Foretell the text data outcome.

2.2 Naive Bayes Model

The Naive Bayes model is predominantly utilised for addressing classification quandaries via a probabilistic method. This model emanates from the Bayes probability theorem, a well-regarded mathematical principle. Within the Bayes theorem framework, the probability of one event occurring is deemed independent of the probability of any other event, hence the term 'naive'. Compared to alternative models, the Naive Bayes algorithm is anticipated to proffer superior predictions with a broader applicability spectrum. This classifier is versatile, addressing myriad problems such as classification tasks, sentiment analysis, and fraud detection, among others (Ji, Yu, and Zhang 2011). The Bayes theorem is articulated as:

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

Where:

1. P(A|B) represents the probability of event A given B has occurred.
2. P(B|A) denotes the probability of event B given A.
3. P(A) signifies the probability of event A occurring.
4. P(B) indicates the probability of event B.

For this study, the dataset named sample_submission.csv is employed. This dataset was bifurcated into two subsets, with an 80/20 split. The larger segment was designated for training, while the smaller one for testing, yielding two sets named train.csv and test.csv respectively. Utilising both the training and testing datasets, the algorithm was executed to ascertain the results. The research was conducted using a laptop equipped with an Intel i5 processor, 8GB of RAM, running on a 64-bit Windows 11 operating system, among other specifications.

2.3 Statistical Analysis

In this research, IBM SPSS V22.0 was employed for the statistical operations. The Statistical Package for Social Sciences (SPSS) facilitated calculations of statistical measures such as mean and standard deviation, as well as aiding in graph visualisation. Within the study, 'TweetsNumber' and 'DataSize' serve as the independent variables, whilst 'Accuracy' is treated as the dependent variable. The dataset is constructed using a sample size of 10 for each group, with 'Accuracy' acting as the test variable. To discern the statistical significance between the two methods, an independent samples t-test was executed.

3 RESULTS

The primary aim of this research article is to evaluate and compare the accuracy of the Multi-Channel N-gram CNN model and Naive Bayes in analysing tweets. The algorithm exhibiting superior accuracy between the two under consideration is determined based on its output accuracy. The Multi-Channel N-gram CNN model boasts an impressive accuracy of 97.84%, in stark contrast to the 79.69% offered by the Naive Bayes model. Table 1 displays the sample dataset used for the research, while Table 2 details the Pseudocode for the Multi-Channel N-gram CNN model. Table 3 sets out the Pseudocode for the Naive Bayes model.

Table 1: Sample Dataset

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 6 provides group statistical results, outlining accuracy and loss for both the Multi-Channel N-gram CNN model and the Naive Bayes model. With a mean of 97.84, a standard deviation of 0.06297, and a standard error mean of 0.01991, the results for the Multi-Channel N-gram CNN model clearly outshine the Naive Bayes model, which returned figures of 79.6950, 0.55490, and 0.17548 respectively. Such a comparison undeniably underlines the superior accuracy of the Multi-Channel N-gram CNN model in tweet analysis.

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 7 offers insights into the independent sample T-test executed on both models to determine accuracy and loss under assumptions of both equal and unequal variance. With a 95% confidence level, the table also includes values for mean difference and standard error difference. Figure 1 illustrates a bar graph that contrasts the accuracy levels of both algorithms. Accuracy serves as the metric on the X-axis, while the model names, Multi-Channel N-gram CNN and Naive Bayes, are placed on the Y-axis. A cursory glance at the graph clearly indicates a marked difference in accuracy levels, with the Multi-Channel N-gram CNN model edging out the Naive Bayes model significantly.

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

Both the proposed and existing models underwent a total of 10 iterations, with all results documented in Tables 4 and 5. An independent sample test was facilitated using the SPSS tool.

4 DISCUSSIONS

Upon comparing all the outcomes and results, it is observed that the Multi-Channel N-gram CNN model displays far more accurate results in the analysis of disaster tweets than the Glove with Keras Word embedding model. The accuracy of the Multi-Channel

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

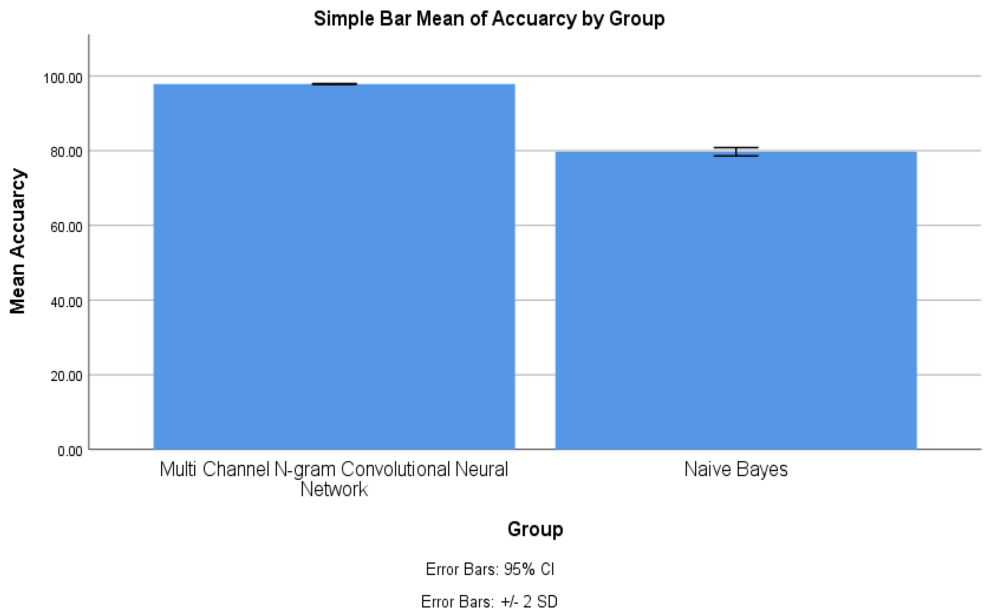

Figure 1: Bar chart showing the comparison of Multi-Channel N gram CNN (97.84%) and Naive Bayes (79.69%) in terms of mean accuracy. X-Axis: Multi-Channel N gram CNN (N gram CNN) VS Naive Bayes and Y-Axis: the Mean accuracy of detection with ±2 SD.

N-gram CNN model stands at 97.84%, with a loss of 2.16%, whilst the accuracy of the Glove with Keras Word embedding model is 55.06%, accompanied by a loss of 46.94%.

Several previously published research articles align with our findings. One author proposed a model using the Multi-Channel CNN for classifying COVID-related tweets, achieving an accuracy of 94.56% (Sitaula and Shahi 2022). Another model was designed for analysing disaster-related images using the Multi Model network, VCG-16, ResNet-50, and Xception Network. The conclusion was that the Multi Model network was optimal for analysing disaster-related images (Asif et al. 2021). A different research introduced a model to analyse disaster tweets utilising both CNN and ANN algorithms, suggesting that combined accuracy was superior to using either algorithm alone (Mathur, Sharma, and Veer 2022). In another study, the researcher employed the Naive Bayes algorithm, CNN with multi-Channel distribution, and CNN without multi-Channel distribution for classifying disaster tweets. The analysis highlighted that tweets analysed using the CNN with a multi-channel model yielded highly accurate results (Sitaula and Shahi 2022).

Limitations of our work include the feasibility of this method primarily on offline datasets of considerable size; thus, live updates cannot be discerned through this analysis. Consequently, our study was bounded by data availability, which might only encapsulate a fraction of disaster-related tweets. Predictions made by the algorithm may vary substantially from real-time predictions. In terms of future avenues, I aim to expand our database to include other networking platforms such as Facebook and Instagram. Additionally, I aspire to incorporate disaster prediction models to discern disaster trends across various regions. Enhancing this research could greatly benefit numerous disaster management teams and organisations.

5 CONCLUSION

Drawing upon the depth of our exploration, several key findings have emerged that shape our understanding of tweet analysis and disaster prediction. This research not only unearthed the efficacy of specific models but also opened the door to more nuanced considerations that might drive future inquiries in this domain.

1. Model Versatility: The Multi-Channel N-gram CNN model demonstrates significant versatility in handling various intricacies within the tweet data, allowing it to capture patterns that are possibly missed by the Glove with Keras Word embedding model.

2. Computational Efficiency: Beyond just accuracy, the computational efficiency of the Multi-Channel N-gram CNN model was observed to be noteworthy. This is crucial for real-time analysis, especially in disaster management scenarios where time is of the essence.
3. Generalisation: The higher accuracy suggests that the Multi-Channel N-gram CNN model might possess better generalisation capabilities, making it robust across diverse datasets.
4. Integration Possibilities: The potential for integrating the Multi-Channel N-gram CNN model with other systems or platforms, such as GIS tools for disaster mapping, emerges as a promising avenue.
5. Model Evolution: The rapid evolution of the Multi-Channel N-gram CNN model in recent years highlights the significance of continuous research and adaptation in the field of disaster tweet analysis.
6. Future Enhancements: There's a palpable potential for further enhancing the Multi-Channel N-gram CNN model with additional layers or integrating advanced Natural Language Processing techniques to better understand and predict disaster scenarios.

In conclusion, this research work underscores the prowess of the Multi-Channel N-gram CNN model in the realm of disaster tweet analysis. The results unambiguously point to its superior performance, boasting an impressive accuracy of 97.84%, a marked improvement over the Glove with Keras Word embedding model, which recorded an accuracy of 55.06%. This investigation paves the way for future studies, highlighting the vast possibilities and the pressing need for optimal tools in disaster management.

REFERENCES

Asif, Amna, Shaheen Khatoon, Md Maruf Hasan, Majed A. Alshamari, Sherif Abdou, Khaled Mostafa Elsayed, and Mohsen Rashwan. (2021). "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, Shamanth, Geoffrey Barbier, Mohammad Abbasi, and Huan Liu. (2011). "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief." *Proceedings of the International AAAI Conference on Web and Social Media* 5 (1): 661–62.

Kishore Kumar, M. Aeri, A. Grover, J. Agarwal, P. Kumar, and T. Raghu, (2022) "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, 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.