CE807 – Assignment 2 - Final Practical Text Analytics and Report

Student ID : 2201912

Abstract

This report focuses on offensive text classification using the HASOC dataset. Two methods, FastText and Convolutional Neural Network (CNN), were employed to identify offensive language. FastText utilizes word embeddings and shallow neural networks for efficient classification, while CNN leverages deep learning to capture intricate features in text. The effectiveness of both methods is compared with respect to the data size of the training dataset and using the validation F1-score.

1 Material

- Google Colab
- Google drive
- Presentation

2 Model Selection (Task 1)

The two types of models I chose for the HASOC text classification are

- FastText
- Convolutional Neural Network (CNN)

2.1 Summary of 2 selected Models

1. Fasttext is a natural language processing (NLP) library developed by Facebook that performs tasks like text classification, word representation, and language identification(Sanoussi et al., 2022). It uses unsupervised learning techniques like skip-gram and continuous bag of words (CBOW) to learn word embeddings and text embeddings, handling out-of-vocabulary words, and training supervised text classifiers with a simple linear model and a hierarchical softmax output layer (Bjørndal, 2019). Fasttext can handle multiple labels per document and weighted labels, and performs k-nearest neighbor search to find the most similar documents or words in a large collection. It provides a command line tool and a Python library for easy usage and offers pre-trained models for various languages and domains(Nitsche and Halbritter, 2019). Fasttext was initially released on November 9, 2015, by Facebook's AI Research (FAIR) lab kumar2023theoretical, and was developed by researchers Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov (Bojanowski et al., 2017).

2. Convolutional Neural Networks (CNNs) are deep learning models that are typically employed for image recognition. Nonetheless, they have also demonstrated potential in tasks involving natural language processing, such as text classification(Young et al., 2018). In the context of offensive language classification,It examines the way text is presented, extract pertinent information, and distinguish between offensive and non-offensive content. Convolutional layers are used in CNNs to capture complex patterns and localised characteristics as they turn incoming text data into a structured representation (Zhang et al., 2018). While fully linked layers link extract features to the top decision-making layer, pooling layers downsample the output. The output of the network is evaluated in comparison to genuine labels, and activation functions like Rectified Linear Units induce nonlinearities. CNNs are useful for named entity recognition, text categorization, and sentiment analysis.

2.2 Critical discussion and justification of model selection

The reason behind selecting FastText and Convolutional Neural Networks (CNNs) as the models to classify HASOC dataset was because of their respective strengths that align well with the specific characteristics of the dataset and classification task.

FastText offers distinct advantages that make it a suitable choice for this task. Given that the dataset involves English text, FastText's ability to generate word embeddings that capture subword information and morphological variations is particularly advantageous. Offensive language in English often comprises variations like slang, misspellings, and abbreviations, which can be challenging to handle using traditional word-based models. FastText's subword embeddings enable it to effectively capture the essence of such terms, ensuring that the model can understand and classify offensive content even in its diverse forms.

Moreover, the dataset consists of short text samples, similar to those found in social media platforms. FastText's efficiency in training and prediction is invaluable when dealing with a substantial volume of short texts. This efficiency is crucial for real-time monitoring and response to offensive content in online platforms, where rapid detection is vital.

For CNNs, their ability to capture local patterns and structures in text aligns with the nature of offensive language detection. Offensive content often exhibits specific word order, phrases, and contextual cues that CNNs can effectively identify and use for classification. This makes CNNs an appropriate choice for this binary classification.

The pipeline of the process from data collection to evaluation is depicted in Table 1

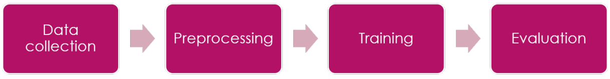

Figure 1: Pipeline flow

3 Design and implementation of Classifiers (Task 2)

The HASOC (Hate Speech and Offensive Content) dataset (Mandl et al., 2019) is made up of 4681 Train dataset, 1171 Valid dataset and 1153 Test dataset. In each dateset, the text are labeled as 'NOT' for post that are not offensive and 'HOF' for text that contains hate,offensive and profane content.Table 1 shows the distribution for the datasets.

Dataset	Total	NOT	HOF
Train	4681	2872	1809
Valid	1171	719	452
Test	1153	865	288

Table 1: Dataset Details

The F1 scores from the validation set of my proposed models FastText and CNN is compared against the state of the art which was gotten by a team called YNU ,they utilized a LSTM approach with ordered neurons and applied an attention mechanism (Mandl et al., 2019). It could be noticed from Table 2 below that LSTM outperformed FastText and CNN with regards to the F1 scores, and FastText outperformed CNN.

Model	F1 Score
FastText	0.64
CNN	0.60
LSTM (Mandl et al., 2019)	0.78

Table 2: Model Performance

4 Data Size Effect (Task 3)

The training dataset was splitted into four sub-set of 25%,50%,75% and 100% using the StratifiedshuffleSplit of Sklearn. Table 3 shows the training dataset size after the split.

Data %	Total	NOT%	HOF%
25%	1171	61.32%	38.68%
50%	2342	61.32%	38.68%
75%	3513	61.32%	38.68%
100%	4682	61.32%	38.68%

Table 3: Comparing model 1 and 2 at 100%

Table 4 compares the output label of the two models AT 100% data size ,it could be noticed that FastText and CNN could give different output labels different from the ground truth (GT). The first example has no offensive text or profane but the GT labelled it as a 'HOF' while the two models classified it as 'NOT'. The fourth text , I am at a loss if it has profane or not, GT and CNN actually labelled 'NOT' while FastText classified it as 'HOF'

Table 5 compares the output label with reference to the data size using FastText as the classifier. The first sentence has no offensive language and all the class size classifies it as ’NOT’.The second text has an offensive language, it was labelled ’NOT’ at 25% but got the right label as the dataset size increased.

Table 6 compares the effect of the datasize in classification of a text as ’HOF’ or ’NOT’. Text with no offensive word always come out right as ’NOT’ but offensive text especially at 25% gives the wrong output labelbut ends up presecting the right class as the data size increases.Surprisingly, a text that has an offensive word predicted a ’NOT’ at 100% dataset. This goes to show the model 2 CNN is prone to false negative output label.

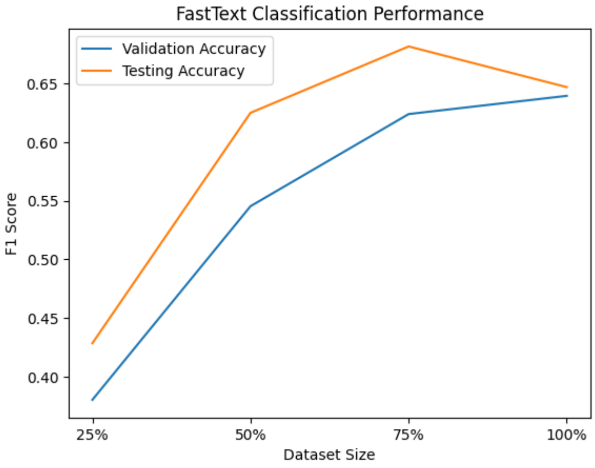

Figure 2: Comparison of Models based on Different Data sizes with FastText.

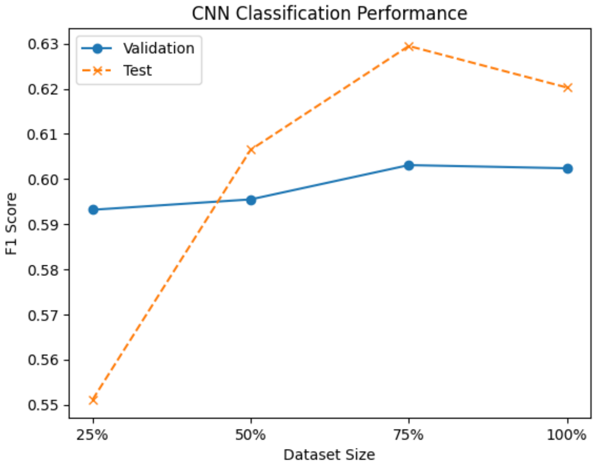

Figure 3: Comparison of Models based on Different Data sizes with CNN

5 Summary(Task 4)

"I conducted text classification using the HASOC dataset, focusing on Task A, which involved categorizing text into offensive or not offensive categories. Employing FastText and Convolutional Neural Networks (CNNs), I explored two distinct approaches for this task. FastText, a powerful word embedding technique, enabled efficient processing and representation of text data, while CNNs captured local and global patterns in the text to discern offensive content. The models were trained using different data size and this showed the size of the train dataset affect the performance of the model performance. The trained models were evaluated using accuracy,recall, precision and F1 score but the F1 score was utilized in comparing with SOA because the data is imbalanced .

5.1 Lesson learnt

Through this assignment,I learnt about the importance of data preprocessing, the impact of training data size on model performance, and the strengths and weaknesses of different model architectures.

References

Bjørndal, S. L. (2019). Exploring pretrained word embeddings for multi-class text classification in norwegian. Master’s thesis, NTNU.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Nitsche, M. and Halbritter, S. (2019). Comparison of neural document classification models. *Hamburg University of Applied Sciences, Department of Computer Science, Hamburg, Germany*.

Sanoussi, M. S. A., Xiaohua, C., Agordzo, G. K., Guindo, M. L., Al Omari, A. M., and Issa, B. M. (2022). Detection of hate speech texts using machine learning algorithm. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0266–0273. IEEE.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Example %	**GT**	**M1(100%)**	**M2(100%)**
"Remember, it is the majority of the English who are the separatists, not the Scots!	HOF	NOT	NOT
At least they told both of you to basically fuck off	HOF	HOF	HOF
’Time for all of us to stand together, to face this Tory belligerence down’ @MaryLouMcDonald	NOT	NOT	HOF
’My prime focus this morning is to speak as loudly and confidently as possible whilst actually spouting the most incoherent stream of shit conceivable.’	NOT	HOF	NOT
Trump story about almost dying in a helicopter crash is a lie: Former Trump Org exec Resist https://www.overleaf.com/project/64479d9fb0a5afcc0bec489fResistTrump FuckTrump	NOT	NOT	NOT

Table 4: Comparison of FastText and CNN using 100% data.

Example %	**GT**	**M1(25%)**	**M1(50%)**	**M1(75%)**	**M1(100%)**
Yes! Boris is only doing what was agreed v early on, a transition period if a deal is struck, is just common sense and was approved all round back in the long lost days of 2017. Having a good Deal, but still wanting to leave next day, simply isn’t necessary or practical	NOT	NOT	NOT	NOT	NOT
Started feeling suspicious when a muslim delivery boy brings food from zomato, after incident of Muslim terrorists trying to poison prasad.	HOF	NOT	NOT	HOF	HOF
Amen girl say that shit over and over again loud and fucking proud!!!!....fucktrump	HOF	NOT	HOF	HOF	HOF
Each of the divorce must be pronounced with a stipulated interval of time between each pronouncement TripleTalaqTruth	NOT	NOT	NOT	NOT	NOT

Table 5: Comparing Model Size: Sample Examples and model output using FastText with different Data Size

Example %	**GT**	**M2(25%)**	**M2(50%)**	**M2(75%)**	**M2(100%)**
Happy Birthday to Sh @nitin-gadkari Ji, God bless you good health and long life sir	NOT	NOT	NOT	NOT	NOT
Jesus Christ Christian News. Illuminati is now changing Bible into gay comedy book: God having bestiality sex, Peter running around naked, King David bowing Muslim way to gay lover Johnathan, ’wink mark’ after incest verse, Jesus wearing girdle, men’s milk https://www.youtube.com/watch pic.twitter.com/SNcTq321bj	HOF	NOT	HOF	HOF	HOF
Madam ji pls also raise your loud voice against the exodus of Kashmiri pandits by your so called ’à¤¡à¤°à¤¾ à¤¹à¥à¤† à¤®à¥à¤¸à¤²à¤®à¤¾à¤¨ ’ who raped,butchered,acid	HOF	HOF	HOF	HOF	NOT
That stupid boy lewismatutu should have simply told their MP to stand down than risky ICC suspension. After all he a a proxy for the owners of this kaundri	HOF	NOT	HOF	HOF	HOF
’I came by boat’ By https://m.facebook.com/Justice-for-refugees-1115243108837143/Â â€¦ Nauru Manus auspol Australia refugees	NOT	NOT	NOT	NOT	NOT

Table 6: Comparing Model Size: Sample Examples and model output using CNN with different Data Size