HCovBi Caps Hate Speech Detection using Convolutional and Bi-Directional Gated Recurrent Unit with Capsule Network

ABSTRACT

Adversaries and anti-social elements have exploited the rapid proliferation of computing technology and online social media in the form of novel security threats, such as fake profiles, hate speech, social bots, and rumors. The hate speech problem on online social networks (OSNs) is also widespread. The existing literature has machine learning approaches for hate speech detection on OSNs. However, the

effectiveness of contextual information at different orientations is understudied. This study presents a novel Convolutional, BiGRU, and Capsule network-based deep learning model, HCovBi-Caps, to classify the hate speech. The proposed model is evaluated over two Twitter-based benchmark datasets – DS1(balanced) and DS2(unbalanced) with the best performance of 0:90, 0:80, and 0:84 respectively considering precision, recall, and f-score over unbalanced dataset. In terms of training and validation accuracy, the proposed model shows the best performance of 0:93 and 0:90, respectively, over the unbalanced dataset. In comparative evaluation, HCovBi-Caps demonstrates a significantly better performance than state of-the-art approaches.

In addition, HCovBi-Caps shows comparatively better performance over the unbalanced dataset. We also investigate the impact of different hyperparameters on the efficacy of HCovBi-Caps to ascertain the selection of their values. We observed that a higher value of routing iterations adversely affects the model performance, whereas a higher value of capsule dimension improves the performance.

**EXISTING SYSTEM**

Warner and Hirschberg [20] used unigram, part of speech, and other template-based features in one of the early approaches to tackle the hate speech problem. The authors further trained the SVMlight model using linear kernel and evaluated it over two datasets from Yahoo and the American Jews Congress websites to classify the hate from non-hate content. In another approach, Kwok and Wand [21] used unigram features and further trained Naive Bayes classifier to segregate the racist tweets from ordinary ones with an accuracy of 76%. They experimentally concluded that bigram, trigram, and sentiment improve model performance. In another approach based on n-gram, Burnap and Williams [22] employed various n-gram features and trained three machine learning models: Bayesian logistic regression, SVM, and voted ensemble classifiers. They further evaluated the trained models over the crawled Twitter dataset and reported that voted ensemble classifiers show the best performance. Djuric et al. [23] utilized paragraph2vec [24] language model for the joint modeling of comments and words collected from the Yahoo Financial website. They further used the trained dense vector representation to learn a logistic regression model to classify the hate comment.

In a popular approach, Waseem and Hovy [25] open-sourced a benchmarked dataset of 16k tweets containing hate speech. The authors further used 1\_4-gram features to train the logistic regression classifier to segregate the hate and ordinary tweets. The best model shows performance with an F1-score of 73:89. They also used location and gender features and gender with n-gram reports the best performance. The various categories of hateful content, such as hate, offensive, abusive, have subtle differences; however, it is understudied. Davidson et al. [18] investigated the difference between hate, abusive, spam, and genuine content and used unigram, bigram, POS tag-based n-grams, Flesch–Kincaid Grade Level, Flesch Reading Ease scores, sentiment score, and various linguistic features to train logistic regression classifier to segregate them. Malmasi and Zampieri [26] presented a similar approach using character and word n-gram features to train linear SVM classifier to classify the hate, offensive, and ordinary contents. They experimented using various feature combinations and reported that character 4-gram shows the best performance.

**Disadvantages**

* The existing literature has no universally accepted definition of hate speech, and even OSNs do not have a consensus.
* An existing system doesn’t Filtering of Twitter-related markers and symbols, such as hashtags, URLs, mentions, and retweets.

Proposed System

\_ Introduce a novel deep neural network model, HCovBi-Caps, by integrating the BiGRU, Convolutional layer and Capsule network to incorporate the contextual information at different orientations for hate speech detection.

\_ Perform the comparative evaluation of HCovBi-Caps over two benchmark datasets to establish its efficacy.

\_ Investigate the impact of different hyper-parameters values on the efficacy of HCovBi-Caps performance to observe the best hyper-parameters values.

**Advantages**

* The HCovBi-Caps applies the convolutional layer over the embedding vector to extract the spatial features. The proposed model uses the one-dimensional convolutional operation because the input embedding vector is a row vector.
* HCovBi-Caps is a best contribution to the collaborative effort going around the world to eradicate the hateful and anti-social content from OSNs.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).