**Doc2Vec**

The word embedding method has been developed to enable the computer to understand texts. The method is developed based on artificial neural networks, and the words are represented as vectors. Quoc Le and Tomas Mikolov developed Doc2Vec to predict a target word by generating a vector that represents the document. The method identifies the vector by not counting the length of the document. Consequently, the method implements two strategies, the Distributed Memory Model of Paragraph Vectors (PV-DM) and the Distributed Bag of Words of Paragraph Vector (PV-DBOW). The PV-DM method accepts each paragraph as a word, and each paragraph is assigned a unique identity, a vector representation. The vectors are started randomly, act as a moving memory, and account for what is missing. In the method, the word vector represents the concept of the word, while the document vector represents the concept of the document. Consequently, the PV-DBOW classifies words using a paragraph vector instead of just guessing the target word. The method consumes little memory and fewer resources because it does not save word vectors.

**Word2Vec**

Word2Vec combines vectors of similar words to establish effectiveness. The method makes a firm estimate about words given a large dataset based on their occurrences in the text. The estimates produce words that are associated with other words in the corpus. For instance, words like kings and queens would be grouped. Additionally, the method performs algebraic operations on word embedding by finding a close approximation of word similarities.

**Neural networks over KNN and Naïve Bayes**

Neural networks are powerful techniques to represent the complex relationships between inputs and outputs. The neural network is a deep machine learning algorithm that takes the input image, assigns importance to various aspects of the picture, and compares the characteristics. Additionally, the preprocessing required in neural networks is lower than KNN and Naïve Bayes. Consequently, the analogy implemented in neural networks is the same connective pattern of the human brain and the organization of the visual cortex. Although KNN requires no training time and neural networks are time-intensive, after the training, neural networks take less time to evaluate data points and estimate research compared to KNN and the naïve Bayes. Additionally, neural networks optimize the procedures using several hyperparameters, which control the network's structure and size.

Text categorization requires faster processing and effectiveness. Neural networks have achieved state of the art in processing and efficiency in very many perspectives compared to KNN. Furthermore, once the neural network is trained, no more training is required to make data predictions because the network will discern the decisions independently. A neural network is considered KNN because when the neural network is trained to handle one task, the same parameters can be used to manage and initialize another similar task. The neural network can facilitate transfer learning without any more training, an aspect that is not present in the KNN and the Naïve Bayes.

**Convolutional Neural Networks**

Convolutional neural networks are a type of neural network used to recognize an image, process the image, and identify its importance, more like interpreting the pixel information. Additionally, they provide a more advanced method of classifying the photos, leveraging principles from linear algebra, identifying patterns in an image, and recognizing object tasks. Convolutional neural networks exhibit superior performance with image, audio signal outputs, and speech. Additionally, it utilizes three layers: the convolutional layer, the pooling layer, and the fully connected layer.

**Ethics of Using Neural Networks**

**Utilitarianism**

When neural networks are used to evaluate statistics and quantitative data, they exhibit advantages and disadvantages. For instance, the algorithm can have significant ethical usage when performing beneficent tasks, like safely directing traffic. The algorithm can effectively identify complex and unintuitive patterns that average human being cannot remember. Additionally, the minority classes are likely to be crashed by the utilitarian philosophy in social situations. For instance, if we use the algorithm to determine whether a bank should grant certain people loans or not, and the race attribute is used as the boundary of the neural networks, the algorithm will make racial judgments. However, it is essential to note that the utilitarian philosophy enables neural networks to perform well. Therefore, it is necessary to set the algorithm with boundaries that do not cause harm.

**Literature Review:**

This project shows various challenges and techniques involved in text categorization. Naïve Bayes and KNN are used here as Naïve Bayes (Pavan, 2021) performs well for categorical input and requires less training data similarly does KNN. Doc2vec model is used as it contributes to better efficiency in categorizing the text so does the author (Akhtar, 2021) explains too.

For even further accuracy, CBOW and Continuous skip-gram methods are also implemented since skip-gram is good for finding rare words and phrases while CBOW works with slightly better accuracy in frequent words (Ria, 2019), thus the combination of these two models yields better accuracy. The training speed can be maximized by using the hybrid model of count-based and window based which is the GloVe model- this study supports the same (Ria, 2019). Further studies show that the use of Bidirectional LSTM is a better approach as there are no restrictions on the flow of data and aids in better prediction (C. Li, G. Zhan, and Z. Li, 2018). In the blog by Cezanne Camacho (Cezanne, 2017) explains the paper of another author (Kim, 2014) that 300 kernels and 100 kernels for each height to cut off the words. This supports the use of CNN and 1 dimensional CNN with max pooling.

The above-mentioned authors have used various methods of deep learning and machine learning techniques and are applied to categorize the news text in this project. These studies show that advanced text categorization is predominantly deep learning techniques with sub domains and are also suitable for large data with multiclass text categorization.

**References:**

Pavan, V. 2021. Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022 [Online] Available at: <https://www.upgrad.com/blog/naive-bayes-explained/>

Dogru, Hasibe & Tilki, Sahra & Jamil, Akhtar & Hameed, Alaa. (2021). Deep Learning-Based Classification of News Texts Using Doc2Vec Model. 91-96. 10.1109/CAIDA51941.2021.9425290

Ria, K. 2019. NLP 101: Word2Vec - skip-gram and CBOW [Online] Available at: <https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314>

Ria, K. 2019. NLP 102: Negative Sampling and GloVe [ Online] Available at: <https://towardsdatascience.com/nlp-101-negative-sampling-and-glove-936c88f3bc68>

C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved Bi-LSTM-CNN," 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893, doi: 10.1109/ITME.2018.00199

Cezanne, C. 2017. CNNs for text classification [Online] Available at: <https://cezannec.github.io/CNN_Text_Classification/>

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification [Online] Available at: <https://www.connectedpapers.com/main/1f6ba0782862ec12a5ec6d7fb608523d55b0c6ba/graph>

A. Bhavani and B. Santhosh Kumar, "A Review of State Art of Text Classification Algorithms," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1484-1490, doi: 10.1109/ICCMC51019.2021.9418262