OSNA INTERMEDIATE REPORT

FAKE NEWS CLASSIFICATION

TEAM MEMBERS

ADITYA SHIVAKUMAR (A20513527) ashivakumar@hawk.iit.edu
SESHA SHAI DATTA KOLLI (A20516330) skolli2@hawk.iit.edu

INTRODUCTION:

Recent years have witnessed an enormous rise in the propagation of disinformation on the internet, which poses an urgent concern considering how swiftly and negatively misleading data can go widespread. The word "fake news" depicts data that has been produced, distorted, or embellished with a view of confusing users. Social networking sites have rendered it simple to create fake news and disseminate it very quickly to a big audience.The simplicity with which information can be produced and shared, the potential to address particular groups with personalized content, and a lack of supervision or influence represent a few of the factors that lead to the propagation of fake news on social media. Additionally, through the promotion of sensationalized or controversial content, social media systems may boost the spread of fake news.

This brings the urgent need for classification of fake news in social media.There are a number of methods that can be used to classify fake news on social media, including, Machine Learning and Natural Language Processing.

In this project, given the title of a fake news article A and the title of a coming news article B, we have been asked to classify B into one of the three categories such as agreed, disagreed and unrelated.

GOALS ACHEIVED:

- For the first step, the training data file has been read. Since the problem description involves splitting the training data into train and validation set, we have split the training data into train and validation set, with a validation split size as 0.2.
- The unnecessary columns (id, tid1, tid2) has been removed from the training and validation data as they would increase the complexity and reduce the model performance.
- Then, special characters are removed the dataset. Removing them would aid in easier analysis of data and model performance.
- All the texts are converted into lower case as this step helps in improving the text normalization, reduces the feature space and noise.
- The stop words are then removed from the text. Stop words are words from the English Language that are typically removed from the text data during processing. These words are regarded as having little significance or being unrelated to the text's content. Examples of stop words are ‘a’, ‘an, ‘the’. Including them makes it even more difficult to analyze the text data as they don’t reduce the feature space.
- The filtered texts are then joined together and the process of tokenization is implemented on the filtered texts. Tokenization is the process of breaking down a text into smaller units called tokens. It breaks unstructured data and natural language text into chunks of information that can be considered as discrete elements. The token occurrences in a document can be used directly as a vector representing that document.
- Once tokenization is implemented, the very next step was to pad the sequences with zeros, so that all the sequences have the same length. Zero padding is an important step because it standardizes all the input data.

- The desired labels (agreed, disagreed, unrelated) were then defined and were mapped to the label column of the training and the validation data.
- To this data we have implemented one-hot encoding. It is a data preprocessing technique that is used to convert categorical data into numerical data.

FUTURE DELIVERABLES:

- The preprocessing of the dataset has been completed. The next step is to perform classification of the text by choosing an appropriate algorithm. Machine learning algorithms like XGBoost, Long- Short Term Memory(LSTM), Passive Agressive classifier etc.
- The future deliverable is to experiment with different algorithm and compare their results and then finally choose the algorithm that provides the best classification results. This step is expected to be completed by by April 20th.
