FAKE NEWS DETECTION PROJECT

PROBLEM STATEMENT, VARIABLE DESCRIPTION & DELIVERABLES

This task is organized by ByteDance, the Platinum Level Sponsor of the conference. ByteDance is a global Internet technology company. Our goal is to build a global content platform that enable people to enjoy various content in various forms. We inform, entertain, and inspire people across language, culture and geography.

One of the challenges which we are facing is to combat different types of fake news. Fake news here refers to all forms of false, inaccurate or misleading information, which now poses a big threat to human civilization.

Fake News Detection Project – Objective & Deliverables

Problem description: At Bytedance, we have created a large-scale database to store existing fake news articles. Any new article must go through a test on the truthfulness of content before being published. We conduct matching between the new article and the articles in the database. Articles identified as containing fake news will be withdrawn after human verification. The accuracy and efficiency of the process, therefore, becomes crucial for us to make the platform safe, reliable, and healthy.

Recommended Project Steps & Guidelines:

1. **Prepare the data:** You're provided with two excel dataset, first level them and merge those two data into one.
2. **Clean the data:** Clean the data, that is, remove the duplicated news articles, remove special characters, numbers etc., correct the spellings of words.
3. **Conduct EDA (Exploratory Data Analysis) on the cleaned Data:** Perform Unigram, Bigram and Trigram analysis on both real and fake news. Create wordcloud on both data based on the subject. Summarize the words and explore the data and then decide your strategy. Make note of any important assumptions that you make.
4. **Convert the text data:** You have to transform each news article into a numerical representation to create a machine learning model. For example, if we have defined our dictionary to have the following words(predictors) {This, is, the, not, awesome, bad, basketball}, and we wanted to transform the text "This is awesome" into a numerical representation, we would have the following numerical representation of that text : (1, 1, 0, 0, 1, 0, 0).
5. **Develop and Validate Samples:** Divide converted data into 2 parts: Development Sample (70%) & Validation Sample (30%). Build your analysis model using the Development Sample, and validate it on the validation sample and then predict on test sample. You can use neural network to create a model.
6. **Improving model accuracy:** Perform various iterations by eliminating or adding the variables(words) to see if the model accuracy is improving or not.

Variable Description:

You're provided with two excel files named as Real_News and Fake_news. In both the files you will find -

Title: The title of the news article.

News_Text: Contains the detail of news article.

Subject: The topic of the article.

Project Deliverables and Submission Dates:

• Identify key traits/features (words, entities, phrases) of news article descriptions which are fraudulent in nature. • Perform Exploratory Data Analysis on the dataset to identify interesting insights from this dataset. • Create a classification model that uses text data features and meta-features and predict which news article are fraudulent or real.	26th March
• Creating PowerPoint Presentation.	29th March
• Creating Video Presentation.	4th April