Fake news analysis modeling

Problem Statement

Fake news refers to news containing deceptive or fabricated contents that are actually groundless; they are intentionally overstated or provide false information. As such, fake news can distort reality and cause social problems, such as self-misdiagnosis of medical issues. Many academic researchers have been collecting data from social and medical media, which are sources of various information flows, and conducting studies to analyze and detect fake news. However, in the case of conventional studies, the features used for analysis are limited, and the consideration for newly added features of social media is lacking. This project module proposes a fake news analysis modeling method to identify a variety of features in terms of spreading information from Twitter, a social media outlet with a good deal of power in terms of spreading information. Furthermore, fake news was analyzed through neural network-based classification modeling by using the preprocessed data and the identified best features in the learning data.

Background

Twitter is a social networking platform that allows its users to send and read micro-blogs of up to 280-characters known as “tweets”. It enables registered users to read and post their tweets through the web, short message service (SMS), and mobile applications. As a global real-time communications platform, Twitter has more than 400 million monthly visitors and 255 million monthly active users around the world. Twitter’s active group of registered members includes World leaders, major athletes, star performers, news organizations, and entertainment outlets. It is currently available in more than 35 languages. A Twitter user can become a follower of a certain user, and based on this feature, a person’s social recognition, status, and influence in a certain area can be checked. When a Twitter user has many followers, it means that the user is highly recognized in the area he/she belongs to. Tweets by an influential Twitter user have a strong influence in terms of information delivery on Twitter because they are highly likely to be read by many people.

Methodology

Step 1: Data collection and dataset preparation

This will involve collection of images from “news-category” and “fake news” database available at Kaggle and preprocessing them, and extracting features.

Step 2: Developing Neural network based classification model

In this step a neural network based MLP classification model for classifying the users into fake or real based on certain parameters is developed.

Step 3: Training and experimentation on datasets

Training and testing is performed on MLP classification model on the “news category” and “fake news” datasets to do the classification accurately.

Step 4: Visualization and analysis

Different visualization and statistical analysis techniques applied to investigate the best features from the collected data.

Experimental Design

Dataset

For the fake news and real news to be used in the analysis, data provided by Kaggle were used. The dataset used is:

https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset.

Kaggle provides global open data for various areas in the CSV or JSON format and provides data for already-confirmed fake news and real news. The collected information consisted of the news article’s headline, writer, date of the Tweet, and real/fake news status, and was stored in the ‘news_info’ table in the database. Twint, a Web-scraping tool, was used to collect the ST that mentioned each news and information of the user who wrote the TW. ST that had mentioned certain news from January 2015 to April 2019 was collected using Twint.

Evaluation Measures

Evaluation is measured in terms of accuracy and different visualization techniques.

Software and Hardware Requirements

Twint, an advanced twitter scraping tool written in Python is used. Scikit-Learns library MLP Classifier is used. Python NLTK (natural language processing toolkit) package is used.