Data Mining Assignment 1

Identify a problem from your own experience that you think would be amenable to data mining. For that problem describe:

Problem: Fake news identification

Fake news is being created and spread at a rapid rate due to technology innovations such as social media. The issue gained attention recently during the 2016 US presidential campaign. During this campaign, the term Fake News was referenced an unprecedented number of times.

1. What the data is.

   Instances of fake news were collected from a list of suspicious websites. Data collection is labor-intensive, as it involves fact-checking for each and every news article. A variety of fact-checking websites perform this analysis on real news. Therefore, one way to collect data on rumors and false news is to take advantage of these websites and to try to automatically scrape information such as the true vs. false headlines.

2. What type of benefit you might hope to get from data mining.

A work is focused on analyzing the fake news detection that is happened on social networks by various data mining techniques to have enough classes of truthfulness. Defining fake news, differentiating it from deceptive news, false news, satire news, misinformation, disinformation, and rumors based on characteristics of data collected.

3. What type of data mining (classification, clustering, etc.) you think would be relevant.

One way is to formulate the fake news detection as a binary classification problem. However, categorize all the news into two classes (fake or real) is difficult because there are cases where the news is partly real and partly fake. To address this problem, add additional classes is a common practice

4. Name one type of data mining that you think would not be relevant, and describe briefly why not.

For this problem, Semi supervised and unsupervised methods are proposed. Fake news detection can also be formulated as a regression task, where the output is a numeric score of truthfulness. Formulating the task in this way can make it less straightforward to do the evaluation. one of the conditions for fake news classifiers to achieve good performances is to have sufficient labeled data. However, to obtain reliable labels requires a lot of time and labor. The task is then formulated as a clustering problem instead of a classification one