

## **Proposal Team 23**

### **Problem Statement and Task Definition:**

The information spread on social media often spans across millions of users and makes an impact immediately. Therefore, the authentication of information has been an important issue given that people make decisions based on the information they collect or receive. We want to build a system that detects fake news through machine learning techniques, which aims to help users obtain correct information under the era of information explosion. This fake news detection system is supervised textual-based, it will output whether the news is true or false given the headline and content paragraphs of the news.

### **Description of the challenges**

Here we talk about the challenges of fake news detection through several perspectives. First, there are many forms which the news can take. The most common are text-based news with a header and body content. Many news now also include images and videos. Yet, text, images and audio are very different in terms of data structure, and they rely on different methods in the domain of AI to tackle. Therefore, in this project we mainly focus on the essential content of news — “text”, where we analyze the authenticity through news header and content.

Second, sometimes there exists a relativity on the judgment of truthiness as it may be based on human's morality, ideology and other perspectives. This higher-level knowledge or connotation poses a challenge for models to acquire.

Last but not least, many fake news related dataset are US-based and are for a certain period. So language can hardly be changed unless other learning techniques are employed. Moreover, certain events are popular over a period, but new events may happen in the future. Thus, the ability to generalize may also be a challenge.

Our task in general can be divided into two parts, analyzing the texts through NLP models where it captures the semantic meaning; then, classifying the processed texts to predict whether it is fake news. Moreover, we have carefully selected a less biased dataset for training in order to address the generalization problem.

### **Input/Output Behavior with Concrete Examples**

#### Input (News header and news content):

Title: Billionaire Donald Trumps Presidential Campaign Is Flat Out Broke

Content: Donald Trump raised just \$29 million for his presidential campaign committee in the first 19 days of October, about half as much as his Democratic rival, putting him at a severe financial disadvantage in the crucial final days of the White House contest, new campaign finance reports filed Thursday night showed. .....

Trump has done what he does best. He talked a big game while bankrupting the Republican Party for his own personal gain. Convincing Republicans to give him their nomination may go down in history as Trumps biggest con of all.

#### Output (Binary: 0/1): 1

Datasets Reference: [WeFake Dataset](#)  
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9395133>

### **Related Works**

我們將從4個方面介紹假新聞探測的方法，最後會聚焦在一種方法上去做研究。

**A. Knowledge-based methods:** 由於假新聞散播不實的資訊誤導大眾，因此最直接的檢測方法就是將新聞內容的陳述比對已知的事實來檢查其真實性，這個過程稱為事實查核 (Fact checking)，需要依靠大量專業知識以及可信度高的資料來源 (例如：維基百科) 去判別新聞的真假。我們可以透過訓練這些資料來建立監督式機器學習模型，來自動檢測新聞文章，但是目前還是高度依賴人類的知識去進行偵測。

**B. Style-based methods:** 將文章風格作為判斷假新聞的基準。文章風格可以從分析具體的文本特徵，例如：詞頻 (Bag-of-words/TF-IDF)、詞性 (Part of speech tagging)、修辭 (Rhetorical Structure Theory, RST)、文法 (Probabilistic Context-Free Grammars, PCFG)，到分析文章的複雜性，像是可讀性、詞彙多樣性 (Type-Token Ratio, TTR) 等。這些特徵反應了作者的寫作特性以及個人態度，藉此可以評估新聞內容的真實程度。

Hu et al. (2019) 比較 RST、LIWC、CNN 模型的準確度，並提出 MCNN-TFW 模型，TFW 是指高頻率單字 (sensitive words) 的權重，可以顯示出單字對真假新聞的重要性，透過計算 TFW 結合 MCNN 提取的語意資訊來進行假新聞偵測。<https://link.springer.com/article/10.1007/s00779-019-01289-y>

Verma et al. (2021) 設計了 WELFake 模型，提取各種語言特徵 (句法、文法、情感、可讀性) 進行詞嵌入 (word embedding)，包括 TF-IDF 和 CV (count vectorizer)，得出比 BERT 和 CNN 更高的準確度。  
<https://ieeexplore.ieee.org/document/9395133>

Wynne and Wint (2019) 研究 word n-grams 和 character n-grams 作為檢測假新聞的特徵，實驗結果發現 character n-grams 比 word n-grams 效果更好。<https://dl.acm.org/doi/10.1145/3366030.3366116>

Horne and Sibel (2017) 除了使用基本語言特徵，還運用可讀性、詞彙多樣性等判斷文章複雜性的特徵去分析真假新聞內容的區別，結果在標題和內容都有明顯差異。假新聞的標題會用更多專有名詞和動詞片語，而內容則缺乏合理的陳述及證據，而文章長度也少於真實性的報導。<https://arxiv.org/abs/1703.09398>

**C. Propagation-based methods:** 近年來社交網站快速發展，造成假新聞傳播的問題，因此藉由分析用戶行為和資訊傳遞模式來預測新聞的真實程度。

**D. Source-based methods:**

## Methodology

**Preprocessing** 前處理：目標為減少文本量，減少差異性，讓文本達到最精簡卻不失其特徵。

方法：包括(1)去除非文本，像是html tag 之，以及標點符號和數字。(2)修正拼寫錯誤。(3)變形還原回單字原樣，像是ing，複數s，過去分詞ed。(4)去掉stopword，像是冠詞，be動詞等對語意不會有太大影響的詞。(5)斷詞，讓文本精簡並易於處理。

## NLP 預訓練模型

Google BERT

BERT，是 Bidirectional Encoder Representations 的縮寫，它從單詞的左右來考慮上下文。這種雙向性有助於模型更好地理解使用單詞的上下文。此外，BERT 的設計目標是能夠進行多任務學習，也就是說，它可以同時執行不同的自然語言處理任務。BERT 也是第一個無監督、深度雙向的自然語言處理模型預訓練系統。它只使用純文本語料庫進行訓練。

模型來源：<https://www.paperswithcode.com/paper/bert-pre-training-of-deep-bidirectional#code>

- 利用以上預處理過的 model，可以省去訓練的成本，直接拿到結果，並且可以更加準確，也避免重複造輪子的行為。
- 將預處理過的文本應用到上述的 model 中，可以拿到對應的結果
- 此方法能直接分好類並預測出結果，可以不需要使用到接下來會討論到的 classification 的方法。

**NLP analyze the texts**: 目標在於將把作為sample的每個文本轉換成 Feature vector, 讓我們能使用分類器將其分類。

bi-gram :

- 透過右方公式, 算出根據前後文的出現機率, 算出每個feature pair對應的feature vector
- 此方法快速, 但前後文的依賴性不高, feature vector 能表達出來前後文的關係很有限

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

Word2Vec - Skip-gram:

- 從一個字詞, 到產生這個字詞的詞向量, 再到預測其他可能會在那個字詞附近出現的詞彙的機率。
- skip-gram 模型的訓練目標是最大化在給定目標詞的情況下預測上下文詞的概率。
- 有點類似圖片做convolution的方式, 可以同時運用到前後文來產生 feature vector, 能有較高的依賴性
- 可以藉由上面的運作方法, 將字串list轉成feature vector。

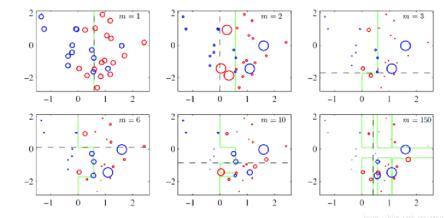
**Classifiers 分類器**: 目標在於根據上一步驟拿到的 feature vector, 以及 database中的vector, 運用監督式學習的模型, 進行假新聞的預測

Decision tree

- 運用feature vector, 使用decision tree 的演算法, 可以訓練出tree。
- 此方法快速且對於只要分兩類的問題, 通常很有效果。
- 如果覺得精準度不足, 可以升級為decision forest, 運用polling的方式, 來增加其精準度。
- 但是對於feature vector的運用稍嫌不足, 且長期依賴性不高。

Adaboost

- 前一個基本分類器分錯的樣本會得到加強, 加權後的全體樣本再次被用來訓練下一個基本分類器。同時, 在每一輪中加入一個新的弱分類器, 直到達到某個預定的足夠小的錯誤率或達到預先指定的最大疊代次數, 達到所謂自適應的分類器。
- 公式:
  - 首先, adaboost為每個分類器給定一個權重值 $\alpha$ , 這些 $\alpha$ 是由錯誤率進行計算而得。
  - 而訓練資料中的每一個樣本, 賦予一個權重, 這些權重構成了向量 D。
- AdaBoost 把多個不同的決策樹用一種非隨機的方式組合起來, 表現出驚人的性能!
  - 把決策樹的準確率大大提高, 可以與 SVM 媲美。
  - 速度快, 且基本不用調參數。
  - 幾乎不 Overfitting。



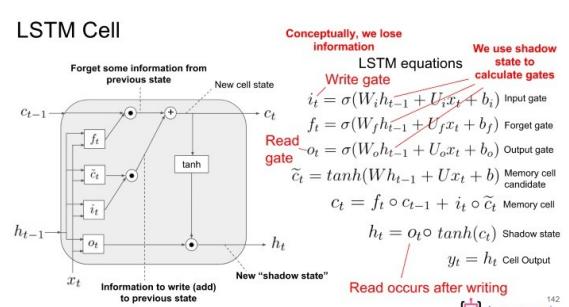
$$\epsilon = \frac{\sum \text{未正确分类的样本的权值}}{\text{样本数目}}$$

$$\alpha = \frac{1}{2} \ln\left(\frac{1-\epsilon}{\epsilon}\right)$$

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{-predict * label * \alpha}}{\text{Sum}(D)}$$

LSTM:

- LSTM 使用記憶來加強當前的決策, 利用三個控制閥來決定記憶的儲存與使用
  - 遺忘閥:如果當前的字句是新主題或以前面字句相反的詞, 那麼, 之前的字句就會這個閥過濾掉, 反之, 可能就會被繼續保留到記憶中。
  - 輸入閥(以  $i_t$  表示):決定當前的輸入及新產生的記憶單元是否加入長期記憶中。
  - 輸出閥:決定當前的字句是否加到輸出, 這個閥也是 Sigmoid 函數, 表示要加入與否。



- 這個方法雖然要使用的資源看似最多，但她可以解決前兩個方法的一個大問題：長期依賴性不高。因為它會自行判斷在甚麼時候捨棄前文。

## Evaluation Metrics

Accuracy:

Precision: 回傳資料的精準度。

$$Precision = \frac{TP}{TP + FP}$$

Recall: 回傳的正確資料佔所有應該被回傳的資料比

$$Recall = \frac{TP}{TP + FN}$$

F1-score: 1:1權衡Precision & Recall

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

MCC(Matthews Correlation Coefficient): 避免資料集偏度過高，導致對分類器的評價失準。

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

loss rate : 算出dataset label與分類的差異

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

## Baselines

- 實作出N-gram + Decision Tree
- F1 score達80分

## Reference

- <https://kknews.cc/code/ey94pkr.html>
- <https://www.tensorflow.org/tutorials/text/word2vec>
- <https://www.itread01.com/content/1547087057.htm>
- [https://docs.google.com/presentation/d/1UHXrKL1oTdgMLoAHHPfMM\\_srDO0BCyJXPmhe4DNhG8/pub?start=false&loop=false&delayms=3000&slide=id.g24de73a70b\\_0\\_0](https://docs.google.com/presentation/d/1UHXrKL1oTdgMLoAHHPfMM_srDO0BCyJXPmhe4DNhG8/pub?start=false&loop=false&delayms=3000&slide=id.g24de73a70b_0_0)
- <https://medium.com/ai%E5%8F%8D%E6%96%97%E5%9F%8E/evaluation-metrics-%E5%88%86%E9%A1%9E%E6%A8%A1%E5%9E%8B-ba17ad826599>
  
- [Automatic Detection of Fake News](#)
- [WELFake: Word Embedding Over Linguistic Features for Fake News Detection](#)
- [Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity](#)
- [Beyond News Contents: The Role of Social Context for Fake News Detection](#)
- [The spread of true and false news online](#)
- [Detecting Pathogenic Social Media Accounts without Content or Network Structure](#)
- [Towards automatic fake news classification - Ghosh - 2018 - Proceedings of the Association for Information Science and Technology - Wiley Online Library](#)
- <http://csjournals.com/IJCSC/PDF12-2/8.%20Jagan.pdf>

Repo link: <https://github.com/dianel0922/Final-Project>