During Sprint one, our client provided the project team with papers to review. Below is a summary with our findings:

dEFEND

Provides explainability for fake news detection using a sentence-comment co-attention sub-network (uses news content and their associated user comments).

1. News content encoding
    a. News content have various levels of linguistic cues, which provide different types of insight in explaining why a certain piece of news is fake.
    b. Learn sentence vectors by using word vectors with attention then sentence encoder
    c. Bidirectional RNN & GRU
2. User comment encoding
    a. User comments can provide useful semantic information
    b. Usually short text
    c. Bidirectional GRU
3. Sentence-comment co-attention
    a. Not all sentences in news articles are fake - many sentences are true, but it only takes one fake sentence for an article to be spreading false information
    b. Takes into account article sentences as well as user comments to better learn weights for them
4. Explainable fake news detection
    a. Concatenate outputs & adds a softmax layer
    b. Minimise cross-entropy loss

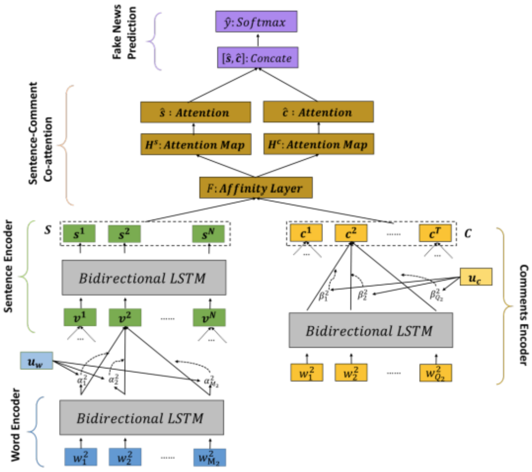


Figure 2: dEFEND Algorithm

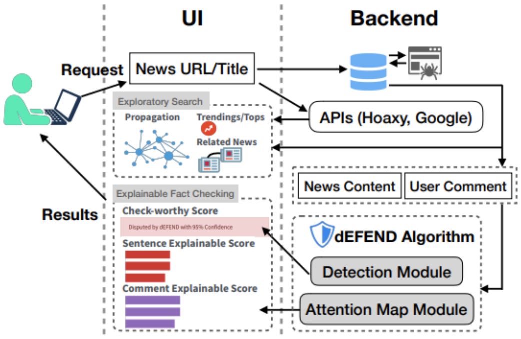


Figure 1: dEFEND System Overview

Quantifying the informativeness of features for fake news detection

Dataset: BuzzFace

Considering 172 features

Unbiased model generation - randomly select a subset of features (up to 20)

Found that for models that were the most accurate, the most common features were: number of shares, reaction count, as well as features that capture political biases and credibility of domain.

XFake

Uses three frameworks to output explanations with visualisation.

1. MIMIC:
    a. Analyses the news attributes, uses deep teacher model to train shallow student model to achieve a combination of good performance and good explanability.
2. ATTN
    a. Semantic analysis of news statements
    b. Pre-trained word embedding, CNN, self-attention to capture global relationships between different words efficiently
3. PERT
    a. Linguistic analysis of news statements
    b. Uses 8 features to train XGBoost classifier, which is then used to make predictions.
    c. Perturbation based method for explanations (observing how much accuracy score changes as a result of a feature being added or removed).

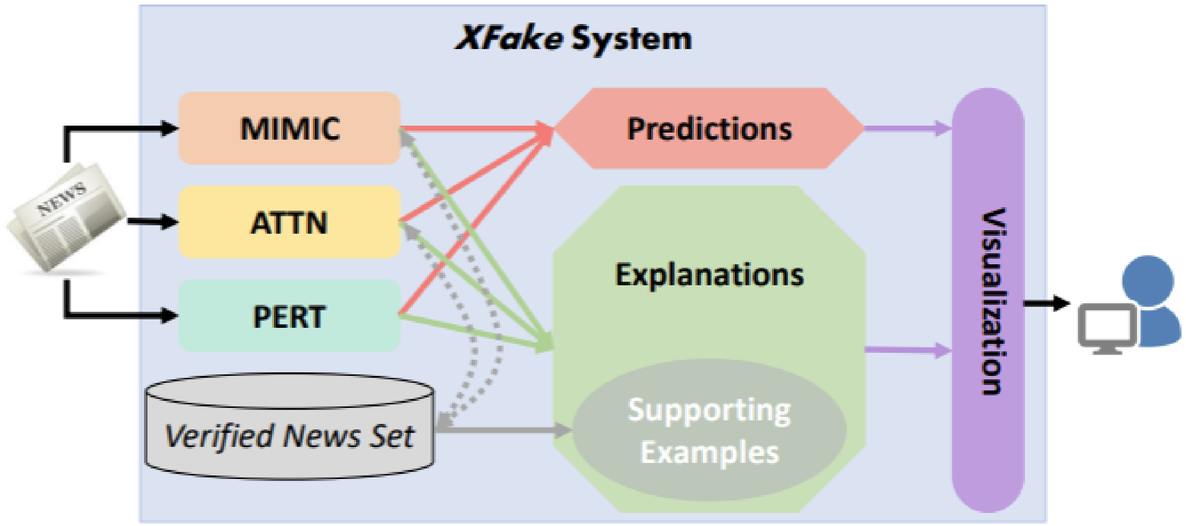


Figure 1: The architecture of XFake system.

.

GCAN (Graph-aware Co-Attention Networks)

Predicts fake news based on the source tweet and its propagation. 90% accuracy

5 components:

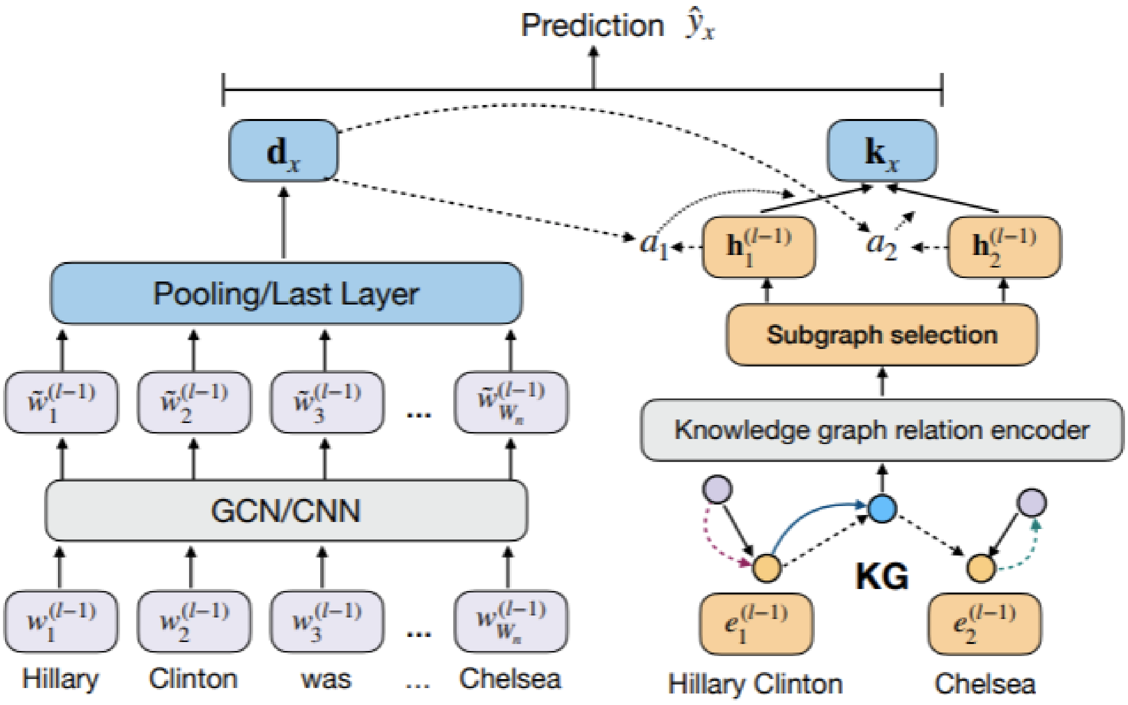
1. User characteristics extraction:
    a. Creates features for how a user participates in online interactions
2. New story encoding:
    a. Create representation of source tweet
3. User propagation representation:
    a. Use GRU and CNN to learn propagation representations with the idea that propagation has different user characteristics depending on how real or fake the news is.
4. Dual co-attention:

    a. Models the mutual influence between source tweet and user propagation, as well as between source tweet and graph-aware representations.
5. Make prediction:

Relational Knowledge

Uses knowledge graph:

- Extracts triples using Stanford NLP tool and uses various techniques to reduce noise.
- Constructs a multi-relational graph, use CompCGN to embed nodes and relations in a relational graph



Fine-Grained Reasoning

Inputs: news article to be verified, online posts regarding the article, users that have published the article

This framework uses two modules:

1. Claim-Evidence Graph Construction
    a. References human process of information storage: extracting the important evidence and removing noise,
    b. Then extract key claims, and associate them with corresponding evidence from above.
2. Graph-Based Fine-Grained Reasoning
    a. Based on Kernel Graph Attention Network to model subtle differences and propagate this information on the graph.

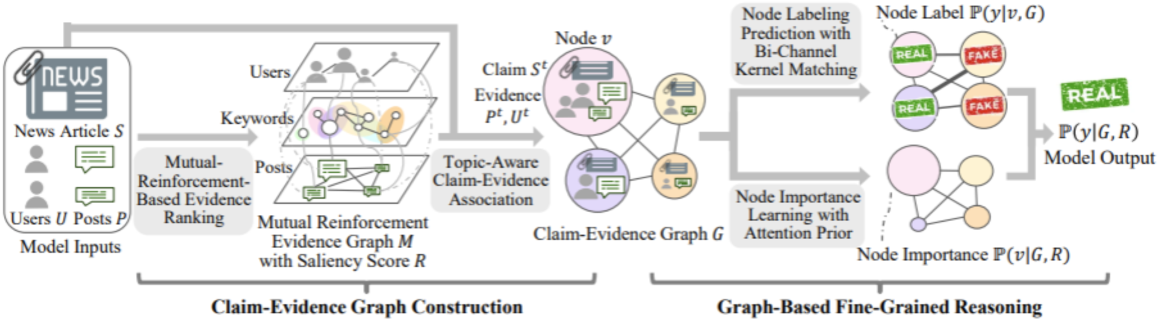

News Article S
Users U Posts P
Model Inputs
Mutual-Reinforcement-Based Evidence Ranking
Users
Keywords
Posts
Mutual Reinforcement Evidence Graph M with Saliency Score R
Topic-Aware Claim-Evidence Association
Node v
Claim S^t
Evidence P^t, U^t
Claim-Evidence Graph G
Node Labeling Prediction with Bi-Channel Kernel Matching
Node Importance Learning with Attention Prior
Node Label $\mathbb{P}(y|v, G)$
REAL
FAKE
REAL
FAKE
Node Importance $\mathbb{P}(v|G, R)$
REAL
$\mathbb{P}(y|G, R)$
Model Output
Claim-Evidence Graph Construction
Graph-Based Fine-Grained Reasoning

Figure 2: Our proposed *FinerFact* framework for fake news detection.

XFlag

LSTM fake news detection model, LRP explanation model

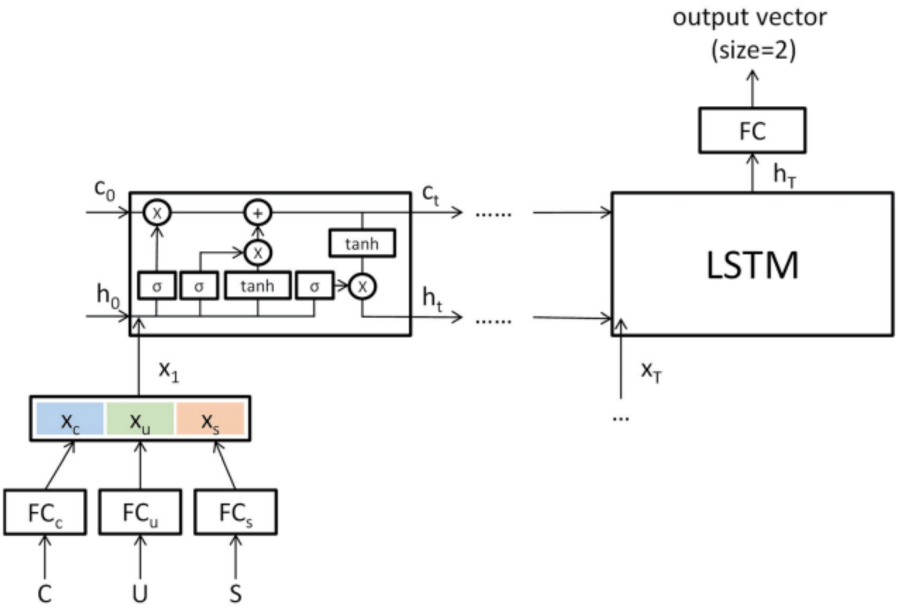


Figure 3. Detection model structure. C represents content, U represents user, and S represents sentiment features. The model output is either a *true* or *fake* news prediction, which is determined by the maximum value in the output vector.