Cross Document Misinformation Detection

CS570 IR Assignment 3: Report

Team Aparichit Retrivers

By

Ashish Giri	224156005	FY MTECH Robotics n AI
Raju Sharma	224156019	FY MTECH Robotics n AI
Sushant Pargaonkar	224156020	FY MTECH Robotics n AI

Problem Statement

Paper:
Given a cluster of topically related news documents, we aim to detect misinformation at both document level and a more fine grained level, event level.

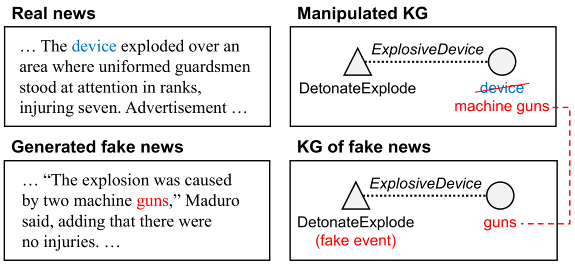
Proposed Novelty:
Given a cluster of documents related to specific topic, task is to generate summary capturing relevant information with the help of GNN representations.

Dataset

Three datasets are derived from “Real News dataset”

- IED - Complex dataset with each complex cluster refers to real world story
- TL17 and Crisis - Timeline summarisation datasets
- Information extraction is performed using OneIE information extraction system which gives us events entities and relations. KG is created based on this data.
- Fake News Generation:
  - Manipulate KG and then regenerate document based on this manipulated KG.
  - For each cluster, we randomly sample 50% real news and replace them with manipulated fake news

Based on the real news and its IE output, we select a high-frequency “DetonateExplode” event and replace its argument entity “device” with “machine guns”. We then generate the fake news from the manipulated KG



… The device exploded over an area where uniformed guardsmen stood at attention in ranks, injuring seven. Advertisement …

… “The explosion was caused by two machine guns,” Maduro said, adding that there were no injuries. …

Example of Fake News Generation

Approach

Task Formulation

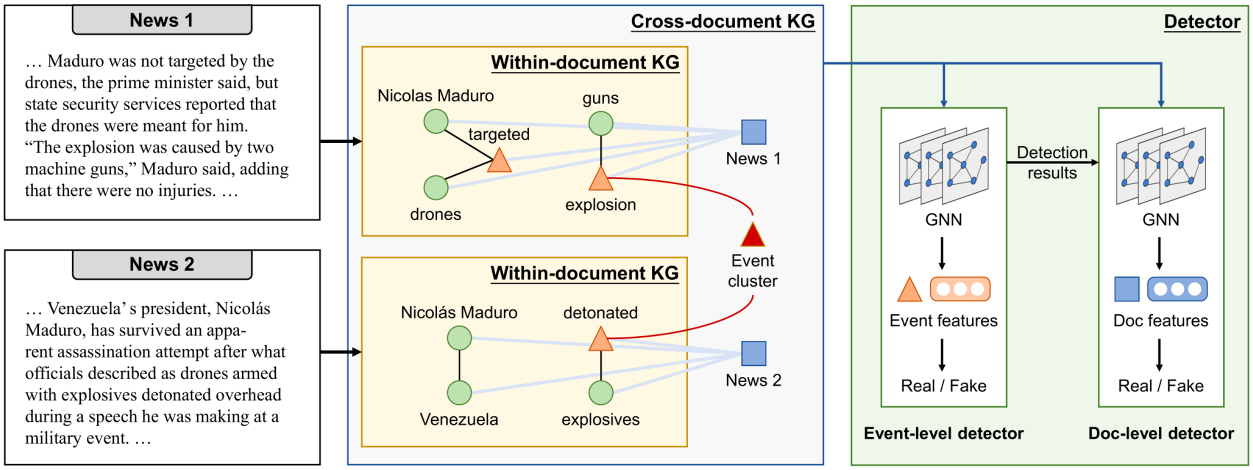
Let S = {d1, d2, … dn} be the document cluster and n = |S| be the size of cluster. Some documents in S are real and others are fake. For each document d in S we extract events E(d) = {e1, e2… em} where m = |E(d)| is number of events in document d. In this extracted event set, some events are real and some events are fake.
Misinformation detection is done at two level, one at document level and at event level. At document level, weather document is fake or real is predicted while event level detection is fine graine task where each event inside the document is considered fake or real

Knowledge Graph Construction

OneIE is a state-of-the-art open-source information extraction system that can automatically identify entities, events, and relations in text. To construct a KG from a text using OneIE, entities, events, and relations must first be identified. Entity linking and coreference resolution are used to merge multiple mentions of the same entity. The resulting KG represents entities and events as nodes and relations as edges between entities, with arguments represented as edges between events and entities.

constructing a cross-document KG involves identifying event clusters from multiple documents, representing them as nodes in the KG, and connecting them with edges to allow for reasoning among cross-document coreferential events. The resulting KG has four types of nodes and five types of edges, which represent the relationships between entities, events, documents, and event clusters.

Knowledge Graph Representation



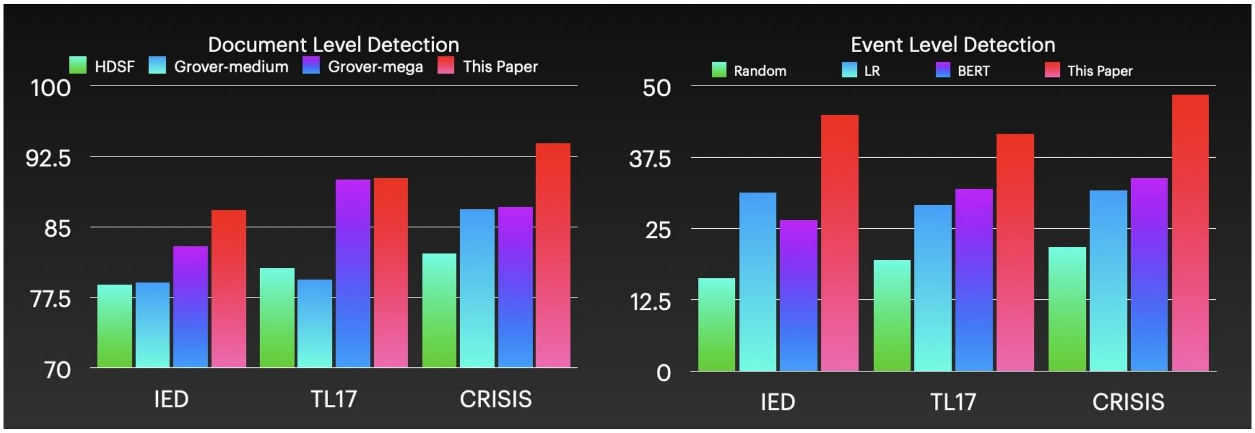
News 1

… Maduro was not targeted by the drones, the prime minister said, but state security services reported that the drones were meant for him. “The explosion was caused by two machine guns,” Maduro said, adding that there were no injuries. …

News 2

… Venezuela’ s president, Nicolás Maduro, has survived an apparent assassination attempt after what officials described as drones armed with explosives detonated overhead during a speech he was making at a military event. …

Results from paper:



Observations of Paper:

Model is heavily dependent on IE system output.

If IE system fails to capture certain events, which are important for classification purpose. Model will fail significantly.

Only one type of mention is taken while generating KG, If we take more type of entities it would help increase accuary.

Limitation and Future work of Paper:

Highly dependent on efficiency of IE system. If IE system fails to capture certain events then results are altered.

Cross document reasoning over more types of information can be conducted to increase the accuracy of the system.

Extension to multimedia news including texts, images, audios and videos by creating Cross Document Multimodal Knowledge Graphs.

Creation of large scale fake news dataset with human written fake news instead of KG manipulated fake news.

Proposed modifications and subtasks:

1. Finding inconsistent information in Medical articles
    a. Thousands of pages claiming different types of remedies for deceases without any scientific background
    b. A good search engine should direct towards most trusted information when queries specially related to health issues.
    c. Taking the same premise of this paper, proposed a way to detect
    d. Setbacks
        i. Generation of dataset - can be generated using websrapping renewed blogs in Medicine
        ii. Generation of Entity, Event and Relations by using OneIE.
        iii. Huge task of data generation. Needs resources.
2. Question Answering
3. Knowledge Graph Based Question Answering (KGQA)
    a. Extracting information from Knowledge graph based on Nodes and Edges
    b. Explored neo4j library for this.
4. News Classification using LLM and GNN representations.
    a. Use both LLM embeddings and GNN embeddings to predict fake or real document as IE system may fail to capture some crucial events which will be captured by LLM
    b. Setback - How to train LLM with multiple inputs for classification task
5. Multidocument Summarization using LLM and GNN.
    a. Discussed below

Multidocument Summarization using LLM and GNN

Multidocument summarization is the task of automatically generating a concise and informative summary from a set of multiple related documents. One way to approach this task is to use a combination of two models: the Language Model for pre-training and the Graph Neural Network for summarization.

The Language Model (LM) is first pre-trained on a large corpus of text to learn the underlying patterns and relationships in language. The pre-trained LM is then fine-tuned on the specific task of summarization, where it learns to generate a summary from a given set of documents.

Next, a Graph Neural Network (GNN) is trained on a graph representation of the documents. The graph is constructed by representing each sentence as a node and connecting nodes that have a semantic relationship. For example, nodes that represent sentences that are similar in meaning or are part of the same event can be connected.

The GNN is trained to predict the importance score of each node in the graph, which represents the relevance of each sentence to the summary. The scores are then used to select the most important sentences for inclusion in the summary.

Finally, the selected sentences are concatenated to generate the final summary.

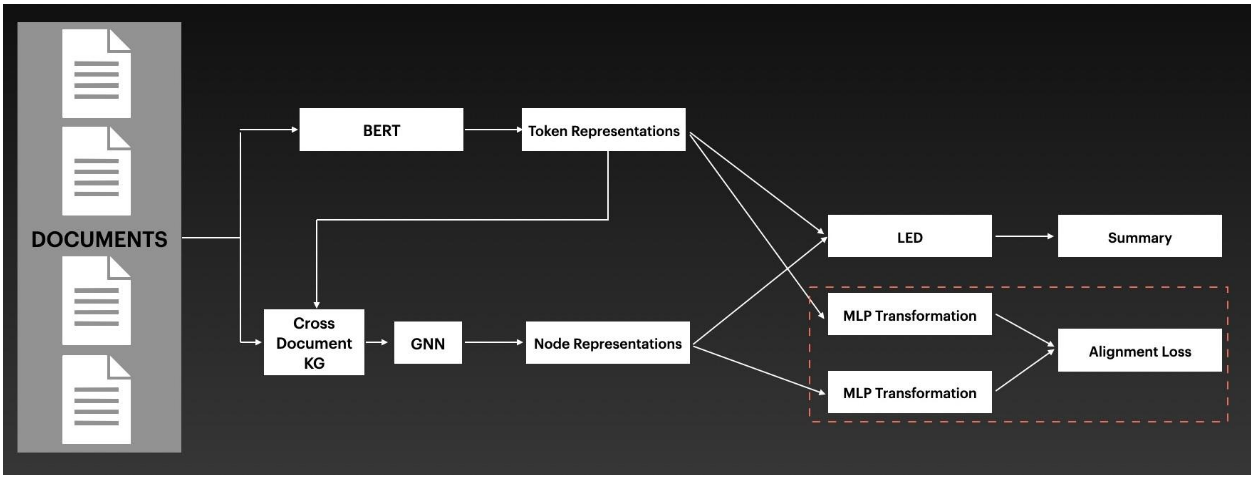
By combining the power of the Language Model and the Graph Neural Network, this approach can generate informative and concise summaries that capture the key information from multiple related documents.

Labelled summary:

We don’t have summaries for this dataset and for summarization we will be needing labelled data to compare out results. Hence we propose to generate summaries from already pretrained model for summarization and cosider it as a ground truth.

Here we have generated summary using **facebook/bart-large-cnn** using hugginface library. We compare our results based on this.

Block Diagram



Results & Observations of proposed task:

<u>True Summary and Generated Summary:</u>

<u>Example 1:</u>

True Label:

International Monetary Fund says Egypt's financial situation is deteriorating. The lending agency won't move ahead with a $4.8 billion loan until receiving updated economic information and reform plans from President Mohammed Morsi's government. Negotiations have dragged on for more than a year for the crucial funding. The money could open the door for more loans and investment, the IMF says. The IMF loan has been delayed by months of negotiations over how Egypt will reduce the huge subsidies. austerity measures lack broad political and social consensus in the highly polarized country where nearly half of more than 85 million people live near or below the poverty line of $2 a day. The 2011 uprising that toppled Egypt's longtime autocratic ruler

Predicted label:
egypt monetary fund says egypt 's financial situation deteriorating lending agency wo n't move ahead 4.8 billion loan receiving updated economic information reform plans president mohammed morsi 's government negotiations dragged year crucial funding money could open door loans investment imf says imf loan delayed months negotiations egypt reduce huge subsidies austerity measures lack broad political social consensus highly polarized country nearly half million people live near poverty line day uprising toppled egypt 's longtime autocratic ruler near investment imf loan delayed months negotiations egypt reduce huge subsidies austerity measures lack broad political social consensus highly polarized country nearly

Example 2:

True Label:
Obama urges Egypt's Mursi to respond to demonstrators. Death toll in clashes between rival protesters since Sunday reaches at least 16. Egypt's armed forces on Monday handed the president a virtual ultimatum to share power. Obama is in Tanzania at the end of an eight-day visit to Africa, including stops in Kenya and Tanzania. of the U.S. President Barack Obama is on a visit to Tanzania. The White House says the president called embattled Egyptian President Mohamed Mursa to urge him to respond. Obama said the political crisis can only be resolved through a political process, the White House said. "Democracy is about more than elections," the statement says. "It is

Predicted label:
mohamed urges egypt 's mursi respond demonstrators death toll clashes rival protesters since sunday reaches least egypt 's armed forces monday handed president virtual ultimatum share power obama tanzania end eight-day visit africa including stops tanzania u.s. president barack obama visit tanzania white house says president called embattled egyptian president mohamed urge respond obama said political crisis resolved political process white house said

Example 3:

True label:
Thousands of Mursi supporters gathered not far from the presidential palace. Some wore had armour, construction hats, shields and clubs. It is thought at least four people have been killed in the protests. The national headquarters of the ruling Muslim Brotherhood was set ablaze. At least seven people, including American student Andrew Pochter, 21, have died in clashes the past week. The protests are expected to continue

for the rest of the week and possibly into next week. For confidential support call the Samaritans in the UK on 08457 90 90 90, visit a local Samaritans branch or click here for details. In the U.S. call the National Suicide Prevention Line on 1-800

Predicted label:
`` egypt mursi supporters gathered far presidential palace wore armour construction hats shields clubs thought least four people killed protests national headquarters ruling muslim brotherhood set ablaze least seven people including american student andrew pochter died clashes past week protests expected continue rest week possibly next week support call uk visit local branch click details u.s. call national suicide line political crisis face details u.s. call national suicide line political crisis face interviewed muslim brotherhood set ablaze least seven people including american student andrew pochter died clashes past week protests expected continue rest week possibly next week support call uk

Observations

predicted label seems to capture the key points of the original text but is not grammatically correct and contains repetitions and omissions.

Conclusions

GNN embeddings captures non linear relationship between entities and events which is helpful for downstream subtask as LLM is not able to capture such information.

Limitations:

1. Need more fine tuning and additional subtask to align LLM embedding and GNN embedding.
2. Output is syntactically not correct
3. Not able to capture Numerical figures

Future Work:

4. Alignment loss between LLM representations and GNN representations
5. Event and Entity recognition subtask based on LLM
6. IE + Summarization model for training!