UNIVERSITY OF SCIENCE - VNUHCM

FACULTY OF INFORMATIC TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE

BIG DATA APPLICATION

PROPOSAL GROUP PROJECT

Group: Allin
Instructors: Prof. PhD. Nguyễn Ngọc Thảo - Teacher Bùi Duy Đăng

FIRST SEMESTER - 2023-2024 SCHOOL YEAR

0 Group information

Group name	Allin
List of members	• 20120090 (Nguyễn Thế Hoàng) • 20120165 (Hồng Nhất Phương) • 20120607 (Lê Hữu Trọng) • 20120609 (Nguyễn Hoàng Trung)

Table 1: Information of group

1 Problem definition

1.1 Overview

Fake news is increasing enormously overtime, becomes a bad source of *Big Data* that people need to identify. As a result, we implement the project "Fake news detection and analyzing" to deal with said problem. Based on news article that we collect from various source, we analyze some prominet features of fake news in various aspect. After that, we also build a model to recognize if a given news is quality or fake, as well as point out which part of news has high chance of being faked.

We also extend the scope of the above problem. That is, most of fake news detection models currently only were trained and work with English news article. Based on collected data, we want our solution gives high quality result on Vietnamese dataset as well. The news may also contains images as its supplementary content, which becomes another challenge that our solution need to deal with.

1.2 Specific tasks of the problem

- **Fake news dection** This is the main part of the problem. There are two sub-tasks that we need to solve:
  - **Binary classification** Basically the model only detects a given news article is "real" or "fake".
  - **Anomaly detection** The model also need to show which part of the news has unusual pattern or characteristics that differ them from credible news.
- **Topic extraction** Retrieving the main content/topic of fake news. We then can analyze to know the topics that have news being faked the most, the frequency and predict shifting in topics in the future.
- **Stance extraction** We retrieve the main targets (person, location, etc.) which are discussed in the news and detect what is the attitude of that news to said targets (Favor, Against, None). This is related to sentiment analysis and can visualize the biased or misleading content.
- **Text analysis** Measure some numerical feature of fake news, such as: text length, lexical diversity (how diversity the way of using vocabulary is), word frequency, readability. These measurement can give us the insight of the property of fake news.

2 Dataset

The followings are the dataset we will collect and use for training fake news detection models and analyzing system:

- **Standadized English dataset** Datasets which have been checked and use by many research before.
  - **Fakeddit** "... a novel multimodal dataset consisting of over 1 million samples from multiple categories of fake news.". The samples are labeled according to 2-way. Each sample contains text and image contents. Most of samples are short news collected from Reddit [1].

    Size: 148 MB (Train set) and 106 GB (Images)
  - **Fake News Corpus** "... an open source dataset composed of millions of news articles mostly scraped from a curated list of 1001 domains". All of them are long article news with text contents [2].

    Size: 9 GB (Whole dataset).
- **Standadized Vietnamese dataset** Same as above but with Vietnamese langague.
  - **VNFD** Contain about 250 long and short articles news from news website and Facebook [3].

    Size: 734 KB (Whote dataset, text version only) - Undefined (Author has not downloaded image of news yet)
- **Self-collected Vietnamese dataset** Because of the scarity of Vietnamese dataset, we decide to crawl suitable Vietnamese article news by ourself. The principle of collecting will be based on the method used in *Fake News Corpus* and *Fakeddit* (for image part). We plan to collect about 1 million news article, as following:
  - **Long article news, credible** Using *scrapy*, *docbao* ([4]) to crawl news from credible source, such as: `vnexpress.net`, `thanhnien.vn`, `quochoi.vn`, etc.
  - **Short article news, credible** We collect short status, notification from Facebook page of above organization.
  - **Long and short news, fake** We plan to list the URL of news website that have high change of giving misleading, uncredible news (such as reactionary page, spam news page, domain that exist in the fake set of *VNFD*).

3 General solution

3.1 Possible solution of each tasks

You only need to search and give name a model/framework/library/tools to sovle for each kind of tasks we have listed in section 1. Describe each of them briefly and how will you plan to use that tool?

- **Fake news detection** The model architecture [5] includes three feature encoding models that used to encode the text, image, and dialogue features, a simple multi-modal embedding, a feedforward network for classification.
    - Multi-Model Feature Encoding
        - **Text Features**: The paper proposed to use RoBERTa and MPNet as the text encoder models
        - **Image Features**: To encode the preprocessed image tensor, ResNet (specifically, ResNet-152) is used in the framework.
        - **Dialogue Features and Summarization Pipeline**: The raw dialogue, which consists of a set of comments, is summarized using the BART summarization pipeline. We plan do not include the comments in the colletected dataset so we may remove this part of the model.
    - Multi-Modal Embedding: After obtaining individual embeddings for each element, they are merged using tensor concatenation. This is followed by a feedforward fully-connected layer with dropout, and the resulting output from this layer is the final multi-modal embedding.
    - Fine-Grained Classification: The final multi-modal embedding is passed through two fully-connected feedforward layers to obtain k-way classification. But in this project we only use binary classification (or two-way).
- **Topic extraction** : Using **NLP** library **Natural Language Toolkit (NLTK)** to extract entity topics, **Latent Dirichlet Allocation (LDA)** technique to retrieve topics in a document collection, and **Gensim** library to implement the algorithm.
    - **Natural Language Toolkit (NLTK)** is a Python library which provides tools for text analysis and natural language processing. It can help analyze extracted content to identify patterns and topics.
    - **Latent Dirichlet Allocation (LDA)** is a popular topic modeling technique for extracting topics from a given corpus.
    - **Gensim** is a Python library for topic modeling, which can help analyze text and retrieve main topics in a set of news and articles
- **Stance extraction** [Huu Trong] Using BERT (Bidirectional Encoder Representations from Transformers) model. To use BERT for stance extraction in fake news, firstly, fine-tuning the pre-trained BERT model on a labeled dataset specific to the task of stance detection. This labeled dataset contain examples of fake news articles or statements along with the corresponding stances (e.g., "support", "deny", "refute",

"neutral"). The fine-tuning process involves updating the parameters of the BERT model using the labeled dataset, so that it learns to accurately predict the stances.

3.2 Architecture

To utilize the overall Big Data architecture, we plan to use *Lambda architecture*. This may help us to set up a standard Big Data architecture, tools, especially when deal with real-time streaming.

We plan to set up that architecture in our project based on this demonstration ([6], that is:

- **Data producing** We choose some news website to get the latest news from there as the real-time information coming to the system.
- **Batch layer** Includes master data set that is immutable, stored in the HDFS. These data can only be appended by the streamed data. We use Hadoop and Spark to compute result for batch views, using the entire of the master data set. As a result, it may takes a great amount of time to compute. We push the computing results to the Cassandra as the batch view.
- **Speed layer** Kafka ingest and manage real-time information and using Spark Streaming to processing streamed data in small time frame. After that we also push them to the Cassandra.
- **Serving layer** Using software framework to get the views from Cassandra and push to the UI (which composes visualization library matplotlib).

4 Weekly planning and expected outcomes

Week	Task	Expected outcome
I	**Collect dataset**	• List the domains to get the data from. • Study the approarch of collecting data from other paper. • Decide where to store data; the format of collected data.
II	**Collect dataset (cont.)**	• Run the collecting tool and push to database.
III	**Data preprocessing**	• Get an overview of the dataset. • Define the requirement of processed data and define the method of running.
IV	**Data preprocessing**	• Implement the defineed preprocessing procedure.
V	**Data analyze and visualization**	• Run the tools/libraries on the preprocessed dataset to collect inforation for the tasks in section 1. • Visualize the result measurement get describe the insight taken from analyzing.
VI	**Train the prediction models**	• Train the fake news detection model on the collected dataset • Train the model for predicting the trend of fake news.
VII	**Making recommendation and setup complete Lambda architecture**	• Utilize Lambda architecture to set up a real-time monitoring system.

Table 2: Weekly planning

References

[1] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection," Nov. 2019. arXiv: 1911.03854. [Online]. Available: http://arxiv.org/abs/1911.03854.

[2] S. Maciej, *FakeNewsCorpus: A dataset of millions of news articles scraped from a curated list of data sources*, 2020. [Online]. Available: https://github.com/several27/FakeNewsCorpus.

[3] H. Thanh, Ninh-pm-se, and T. C. Vi, *VFND/VFND-vietnamese-fake-news-datasets: Tp hp các bài báo và các bài vi´t trên mạng xã hi ti´ng Vi.t phân loại 2 nhãn Tht & Gia (254 bài) và công cụ h˜ tr.*, May 2022. DOI: 10.5281/zenodo.6590948. [Online]. Available: https://doi.org/10.5281/zenodo.6590948.

[4] L. Đng Hai, *docbao: Công cụ quét và phân tích t` khoá các trang báo mạng Vi.t Nam*, 2021. [Online]. Available: https://github.com/hailoc12/docbao.

[5] F. Rahman, "Multi-Model Fine-Grained Fake News Detection with Dialouge Summarization," Yale University, Tech. Rep., 2021. [Online]. Available: https://github.com/faiazrahman/Multimodal-Fake-News-Detection/blob/main/paper.pdf.

[6] A. Souza, *Big-data-pipeline-lambda-arch*. [Online]. Available: https://github.com/apssouza22/big-data-pipeline-lambda-arch.