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.

1.3 Possible solution of each tasks

1.3.1 Fake news detection

The model architecture [1] 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).

1.3.2 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

1.3.3 Stance extraction

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.

2 Dataset Relating Tasks

2.1 Collection

Beside the existing dataset we have found and described in the last proposal, we have collected more Vietnamese dataset by ourselves:

- **Long article news, credible** We have used Selenium to crawl from: `vnexpress.net`, `vov.vn`, tuoitre.vn, vietnamnet.vn. For each these news website, we collected about 2500 recent news belong to various topics. In total, we have collected 10000 news article, and each article has a feature picture.
- **Short article news, credible** We decide do not collect this kind of dataset anymore.
- **Long and short news, fake** We have only collected about 500 news articles from `voatiengviet.com`.

2.2 Data preprocessing

Data preprocessing is an essential step in preparing the collected Vietnamese news dataset for analysis and modeling. Techniques that will be applied to collected dataset:

1. Tokenizations:
    - Tokenization involves splitting the text into individual words or tokens. This step helps in breaking down the text into smaller units for further analysis and processing.
    - Python's built-in split() function can be used to split the text into tokens based on whitespace.
    - The nltk.tokenize module from the Natural Language Toolkit (NLTK) provides various tokenization methods, such as word tokenization (word_tokenize) or regular expression tokenization (RegexpTokenizer)
2. Text cleaning:
    - The raw news articles may contain irrelevant information, such as HTML tags, special characters, punctuation marks, and numbers. These elements need to be removed or replaced to ensure cleaner text data.
    - Python's re module allows to use regular expressions for pattern matching and replacement, which can be useful for removing HTML tags, special characters, or numbers from the text.
    - Beautiful Soup (beautifulsoup4) is a Python library that facilitates parsing and manipulating HTML or XML documents. It can be used to extract text content from HTML tags.
3. Stop Word Removal:
    - Vietnamese language has a set of commonly used words that do not carry significant meaning, such as articles, prepositions, and conjunctions. These stop

words can be removed from the text to reduce noise and improve computational efficiency.

- nltk.corpus from NLTK provides stop word lists for different languages, including Vietnamese. Use the Vietnamese stop word list (stopwords.words('vietnamese')) for removing stop words from the text.

4. Lowercasing:

    - Convert all the text to lowercase to ensure that words with the same meaning but different cases are treated as the same word. This step helps to prevent redundancy and improve the accuracy of subsequent tasks like text classification.
    - Python's lower() method can be applied to convert the text to lowercase.

5. Lemmatization and Stemming:

    - Reduce words to their base or root form to consolidate similar words. Lemmatization aims to convert words to their dictionary form, while stemming simplifies words by removing prefixes or suffixes. This step helps in reducing the dimensionality of the data and improving the efficiency of subsequent analysis.
    - SpaCy is a popular Python library for natural language processing. It provides lemmatization and stemming capabilities. Use the Vietnamese language model (spacy.load('xx_ent_wiki_sm')) in SpaCy for lemmatization.
    - The nltk.stem module in NLTK provides various stemmers, such as the Porter stemmer (PorterStemmer) or Snowball stemmer (SnowballStemmer).

3 Obstacle left

1. Team members currently do not have enough relevant knowledge to solve the problem and need to spend more time to learn more thoroughly.
2. Collecting data becomes cumbersome because it is collected from many sources and the structures of each source are different.
3. Collecting and processing data is complicated when having to process text data in both English and Vietnamese, as well as processing image data.
4. Determining which part of the news has unusual characteristics can cause complications or conflicts when combined with the original binary (real, fake) classification model.
5. Large models and heavy calculations make the team's resources insufficient

4 Weekly planning and expected outcomes

We adjust the proposed plan as following:

Week	Task	Expected outcome
III	**Data preprocessing**	• Implement the defineed preprocessing procedure.
IV	**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.
V	**Train the prediction models**	• Train the fake news detection model on the collected dataset • Train the model for predicting the trend of fake news.
VI	**Building Real-time architecture (Optional)**	• Utilize Kafka architecture to set up a real-time monitoring system.

Table 2: Weekly planning - Changed version

References

[1] 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.