Fake news detection

Giảng viên hướng dẫn: TS.Vũ Huy Thế

Phạm Ngọc Đông

Đoàn Quang Khải

Nguyễn Văn Sơn

Nguyễn Thế Hiển

Nội dung

1. Giới thiệu
2. Dữ liệu
3. Models
4. Kết quả thực nghiệm

1. Giới thiệu

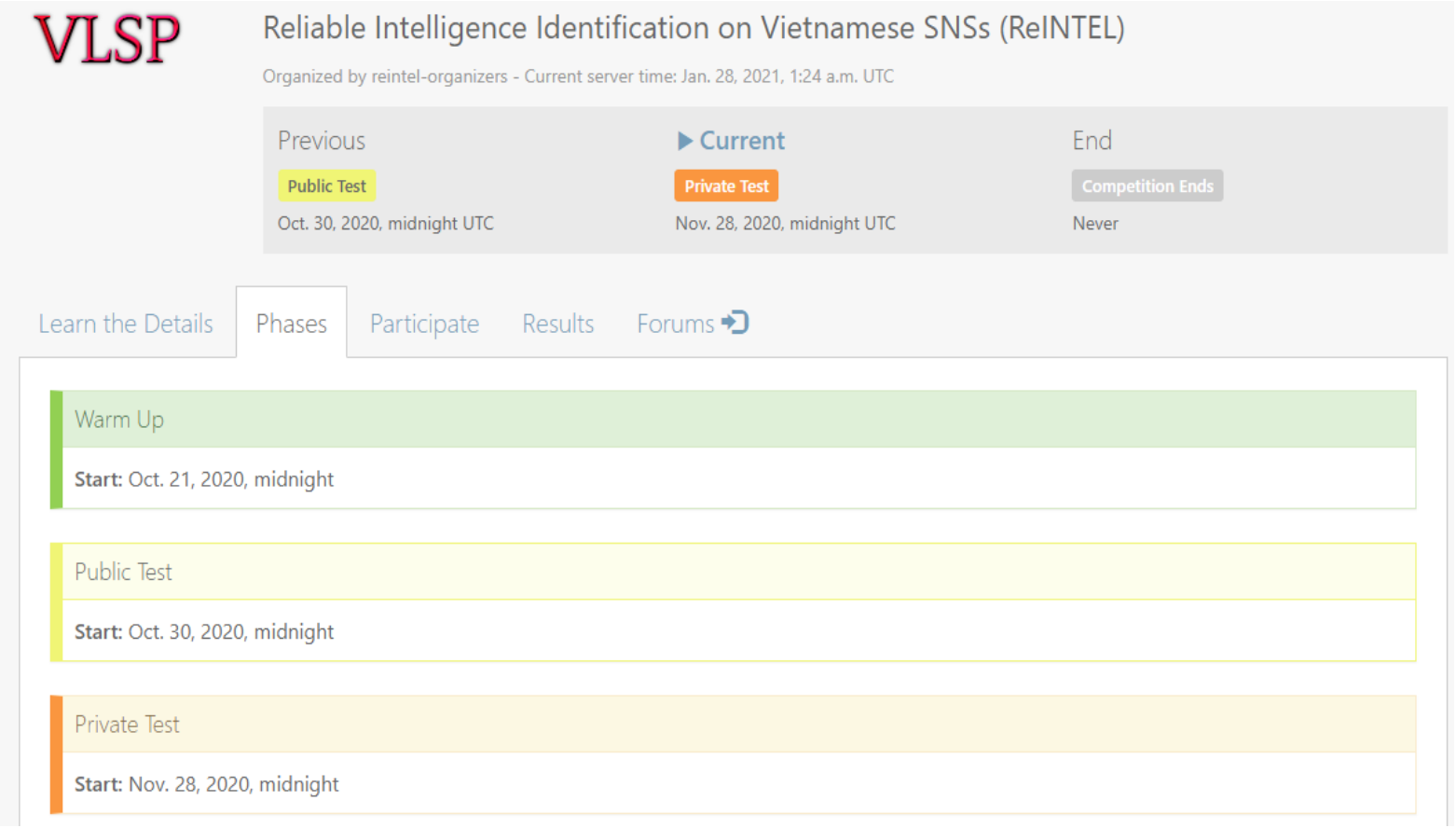
- Tin giả là những tin sai sự thật
- Mục đích:
  + Câu view, like, share.
  + Hướng vào đối tượng cụ thể.

- Đa phần các tin giả gây ảnh hưởng tiêu cực

- Khó để kiểm soát tin giả đặc biệt các tin trên mạng xã hội.
- Tin giả khó để kiểm chứng so với tin thật
- Người đọc tin trên mạng xã hội thường không kiểm chứng tinh đúng đắn của tin, chia sẻ, like theo đám đông.

2. Dữ liệu

❖VLSP 2020 competition



Timeline of VLSP 2020 competition

This challenge aims to identify a piece of information shared on social network (SNSs), is reliable or unreliable.

2. Dữ liệu

❖ EDA data

	id	user_name	post_message	timestamp_post	num_like_post	num_comment_post	num_share_post	label
0	1	389c669730cb6c54314a46be785cea42	THĂNG CẤP BẬC HÀM ĐỐI VỚI 2 CÁN BỘ, CHIẾN SỸ H...	1585945439	19477	378	173.0	0
1	2	775baa6d037b6d359b229a656eaeaf08	<URL>	1588939166.0	11	5	3	0
2	3	b9f3394d2aff86d85974f5040c401f08	TƯ VẤN MÙA THI: Cách nộp hồ sơ để trúng tuyển ...	1591405213	48	5	19.0	0
3	4	808e278b22ec6b96f2faf7447d10cd8e	Cơ quan Cạnh tranh và Thị trường Anh quyết địn...	1592023613	3	0	0.0	0
4	5	f81bdd6d8be4c5f64bb664214e47aced	Thêm 7 ca tại Quảng Nam liên quan đến hành khá...	1583737358	775	0	54.0	0
...	...	...	...	...	...	...	...	...
4367	4368	20933f35ef5d22b4d8193cc269c8ff1e	BÀ MẸ VIỆT NAM ANH HÙNG 95 TUỔI MAY KHẨU TRANG...	1584795126.0	5800	1300	12000	0
4368	4369	a117312f796a22e364b8e241b8cb91eb	Nguồn cung khan kiếm nhưng nhu cầu cao tạo áp ...	1590645643	21	1	NaN	0
4369	4370	547ba1b4f95ec07f2cdada24a6eec693	Lời cảnh tỉnh cho các thanh niên dân TỔ...tốc ...	1589774421.0	3	1	NaN	1
4370	4371	acb4a36d6247a0c89dac880725b2b3a0	Đến bây giờ mới biết chỉ cần học lái xe hạng B...	1589551407.0	144	38	87	1
4371	4372	3deabd01107da8ae2a29ca03483714d1	Tư lệnh ngành cấm bay với phi công Pakistan, r...	1593319282.0	24	9	NaN	0

4372 rows × 8 columns

Overview public train data

2. Dữ liệu

❖ EDA data

- **INPUT:** id, uid, text, timestamp, nb_likes, nb_comments, nb_shares, image_links.
- **OUTPUT:** label

Data Format

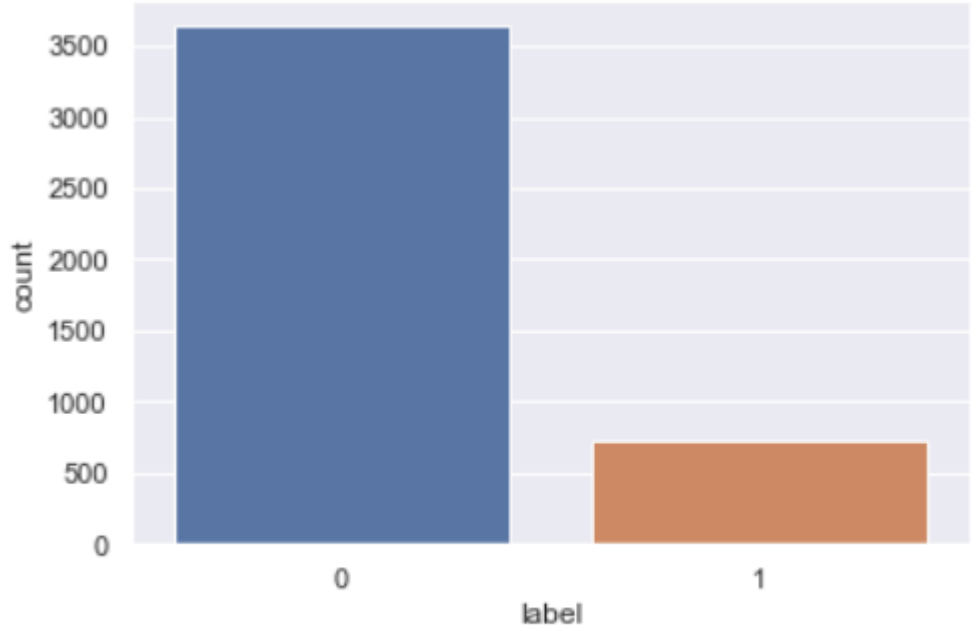
Each instance includes 6 main attributes with/without a binary target label as follows:

- id: unique id for a news post on SNSs
- uid: the anonymized id of the owner
- text: the text content of the news
- timestamp: the time when the news is posted
- image_links: image urls associated with the news
- nb_likes: the number of likes that the news is received
- nb_comments: the number of comment that the news is received
- nb_shares: the number of shares that the news is received
- label: a manually annotated label which marks the news as potentially unreliable
  - 1: unreliable
  - 0: reliable

2. Dữ liệu

❖ EDA data – các vấn đề của dữ liệu

- Mất cân bằng dữ liệu




```
Percent of 0 label: 83.21%
Percent of 1 label: 16.79%
```

- Missing data - NaN

timestamp_post	num_like_post	num_comment_post	num_share_post	label
NaN	4700	549	3300	0
NaN	9400	939	3400	1
NaN	3700	277	4700	0
NaN	10000	657	4700	0
NaN	23000	1300	4800	0
...	...	...	...	...
NaN	3000	974	4300	0
NaN	1100	109	2200	0
NaN	18000	1300	26000	0
NaN	4400	976	4400	1
NaN	4800	436	3100	1

2. Dữ liệu

❖ EDA data – các vấn đề của dữ liệu

- unknown

num_like_post	num_comment_post	num_share_post	label
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1
unknown	unknown	unknown	1

- Trùng lặp nội dung nhưng nhãn khác nhau

```
                                                post_message num_like_post  \
506    "Con virus corona này là một thảm họa tương đư...            48
3516   "Con virus corona này là một thảm họa tương đư...           214

      num_comment_post num_share_post  label
506                  6             74      0
3516                19            474      1
**********************************************************************
```

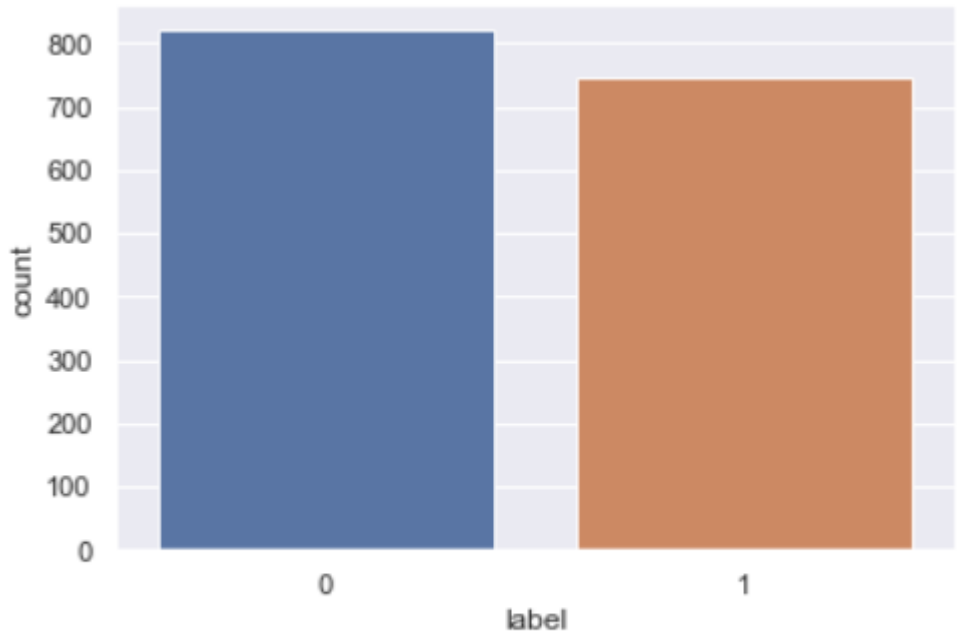
- URL của post_message

	post_message	num_like_post	num_comment_post	num_share_post	label
1	<URL>	11	5	3	0
150	<URL>	87	4	1	0
175	<URL>	NaN	0	NaN	0
178	<URL>	3	2	NaN	0
454	<URL>	4	0	NaN	0
486	<URL>	NaN	0	NaN	0
643	<URL>	28	6	1	0
679	<URL>	1	0	NaN	0

2. Dữ liệu

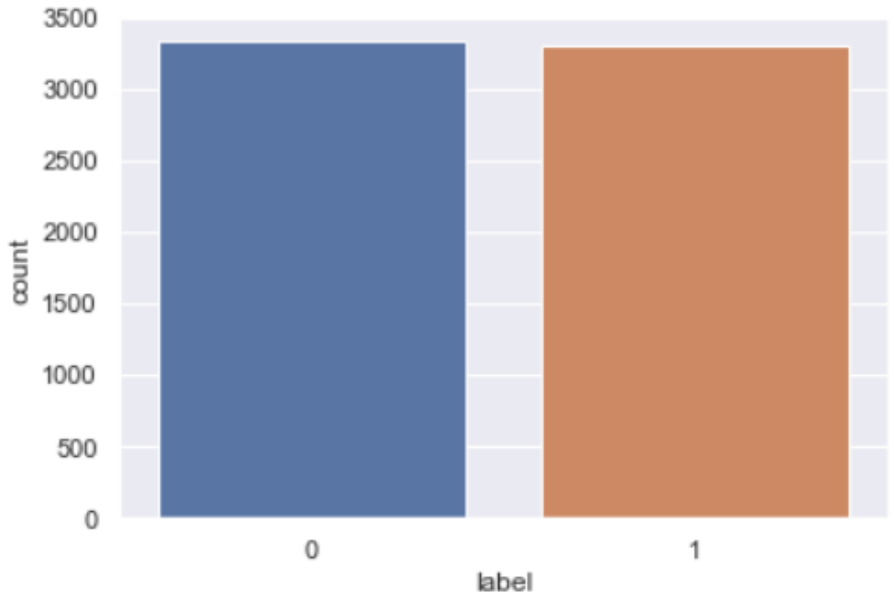
❖ Dreprocessing data - Mất cân bằng dữ liệu

✓ Under sampling




```
Percent of 0 label: 52.36%
Percent of 1 label: 47.64%
```

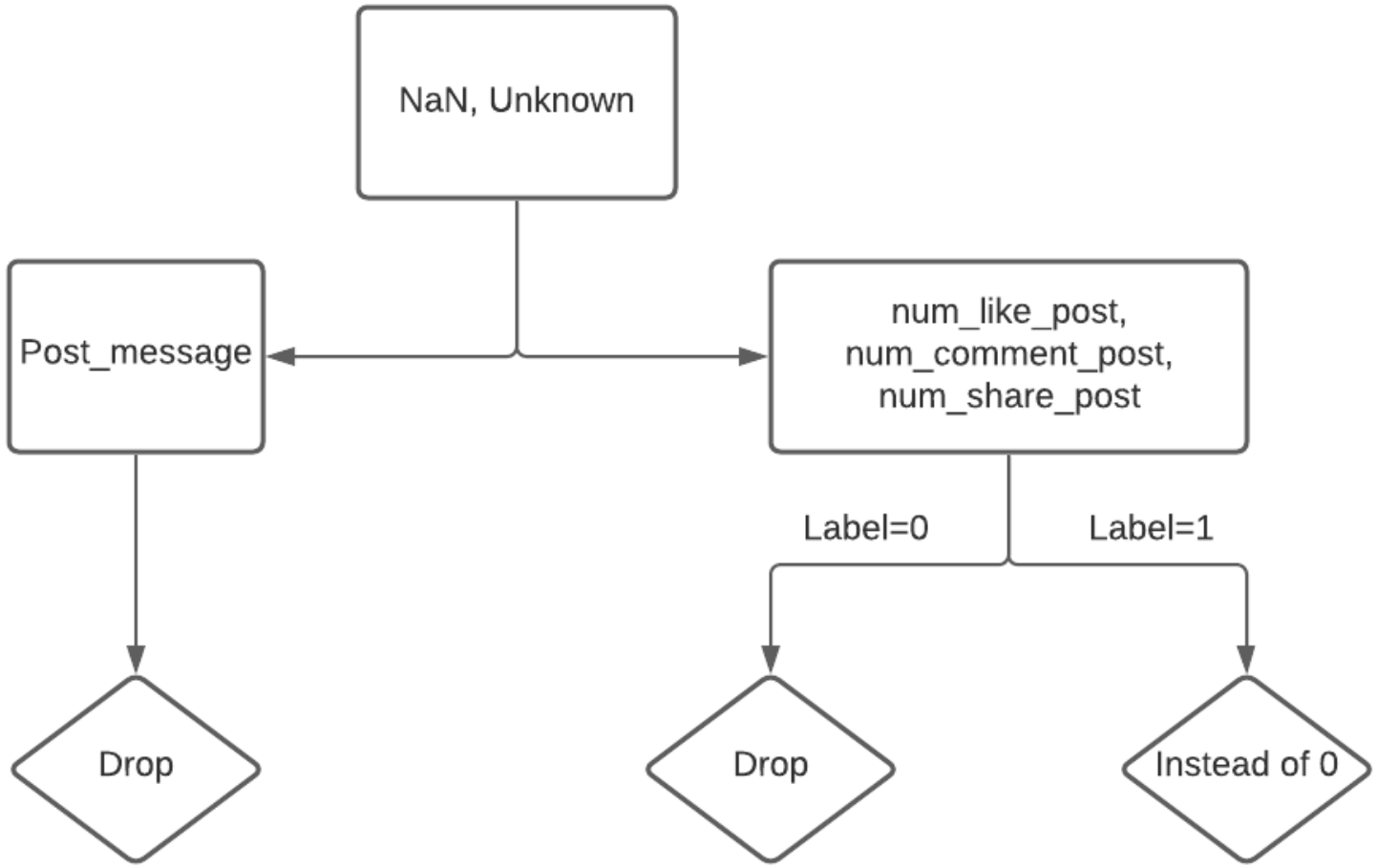
✓ Over sampling




```
Percent of 1 label: 49.74%
Percent of 0 label: 50.26%
```

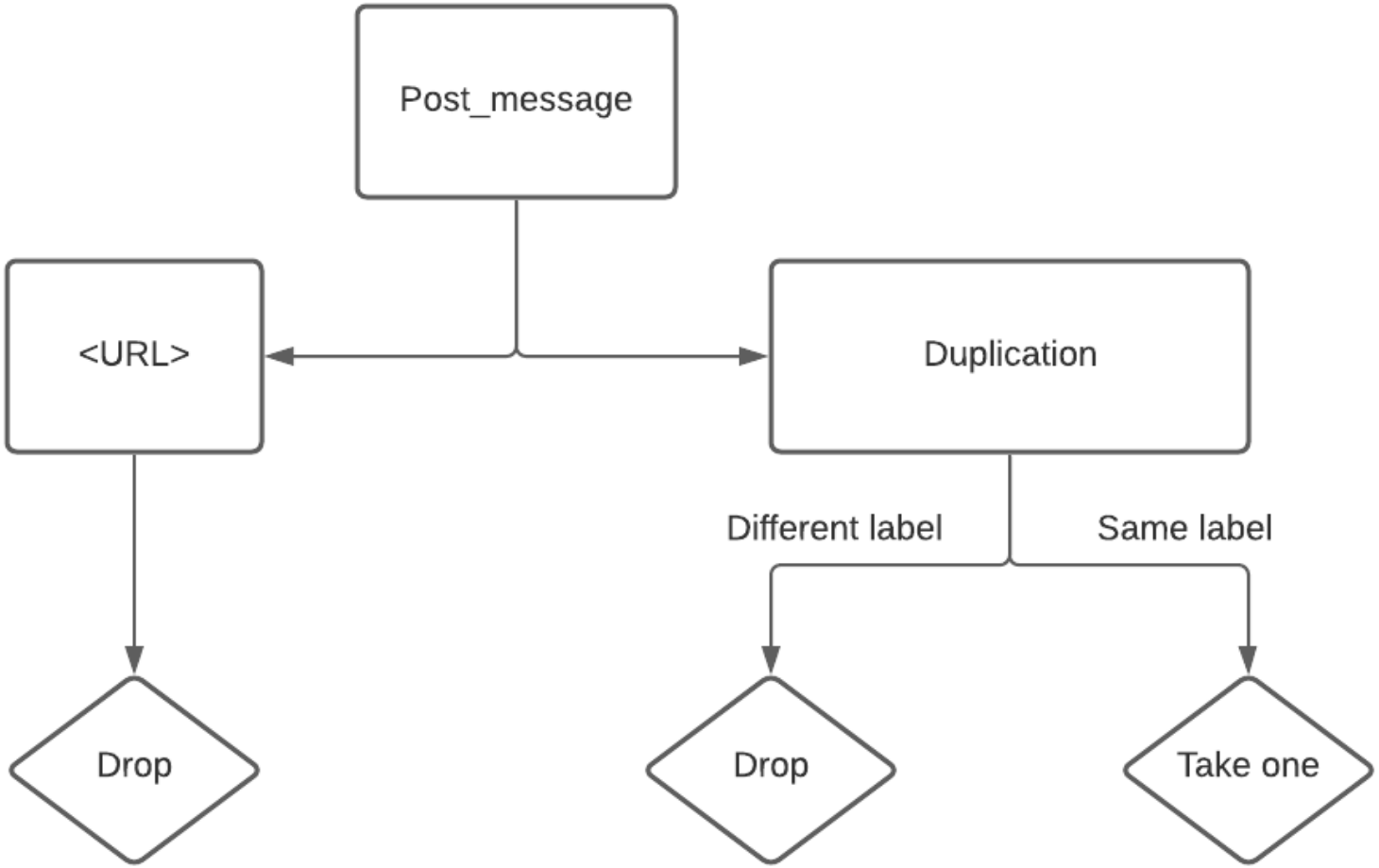
2. Dữ liệu

❖ **Preprocessing data - Missing data – NaN, unknown**



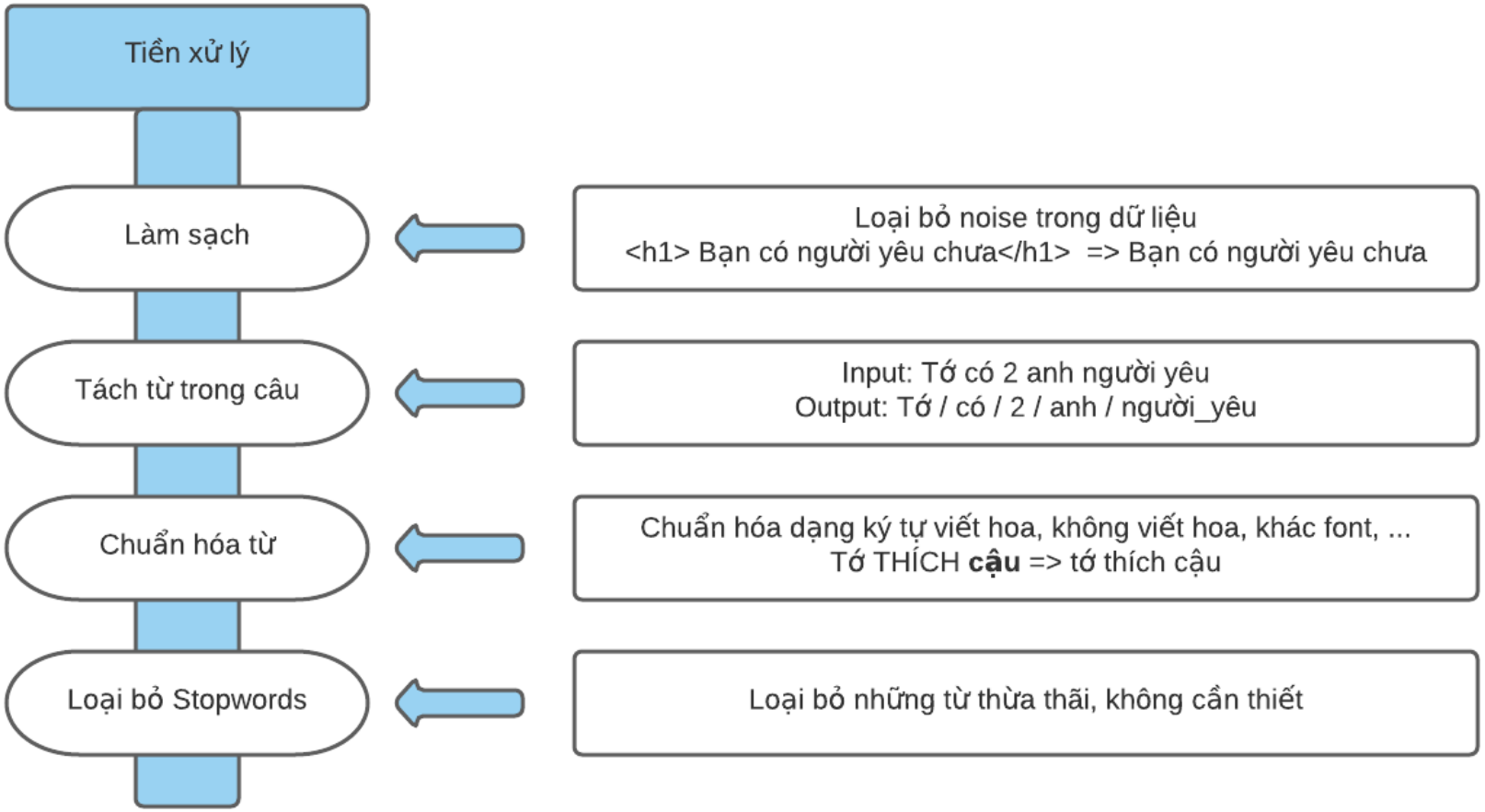
2. Dữ liệu

❖ **Preprocessing - Duplicate and URL post_message**



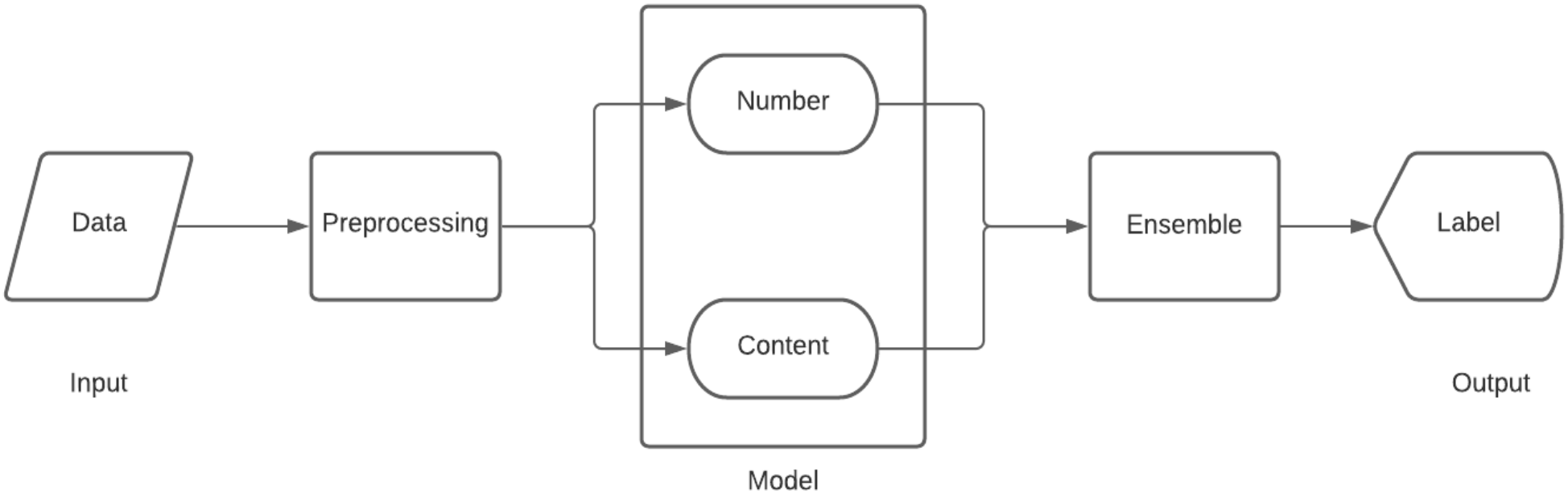
2. Dữ liệu

❖ Preprocessing - Text of post_message



3. Model

❖ Work flow



- ❑ **Content model:** input: post_message, ouput: label or vector
- ❑ **Number model:** user_name, num_like_post, num_share_post, num_comment_post

4. Kết quả thực nghiệm

❖ **Content models**

	Accuracy	F1-score
TF-IDF + RandomForest	0.70	0.73
TF-IDF + SVM	0.85	0.86
TF-IDF + Logistic Regression	0.80	0.80
Pretrain PhoBERT + RobertaForSequenceClassification	0.89	0.89
Pretrain PhoBERT + TF-IDF + RobertaForSequenceClassification	0.90	0.90

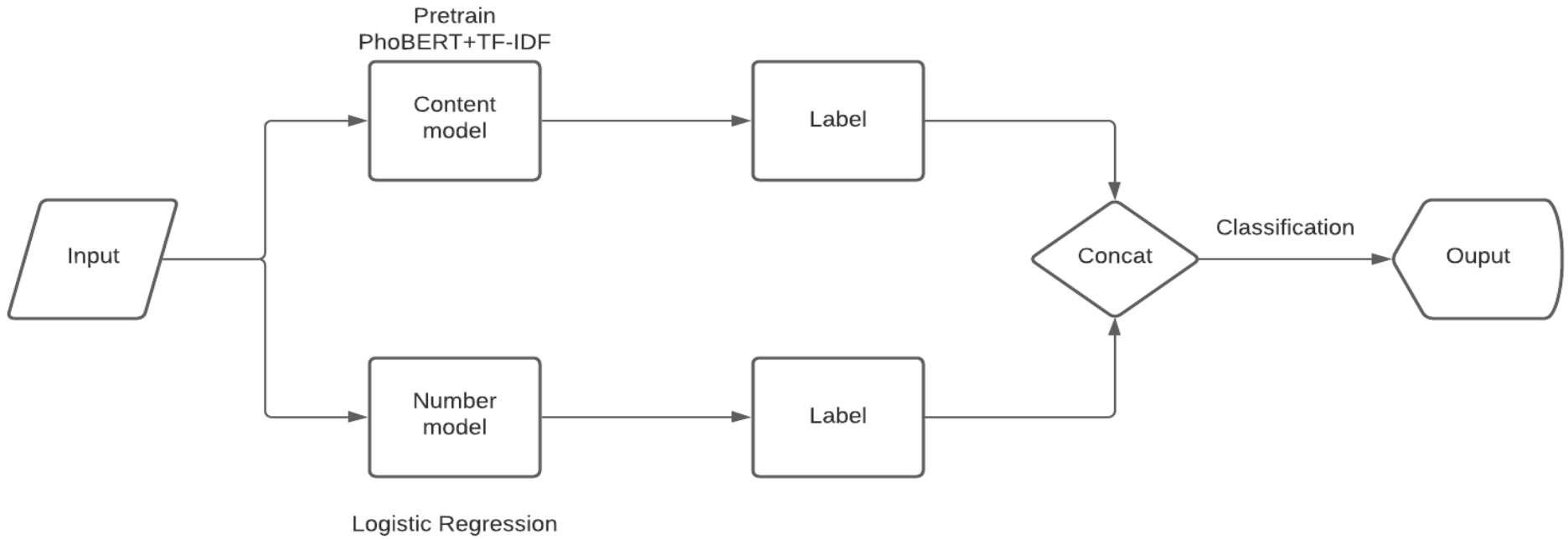
4. Kết quả thực nghiệm

❖ **Number models**

	Accuracy	F1-score
KNN	0.65	0.69
SVM	0.82	0.78
DecisionTree	0.65	0.69
Logistic Regression	0.82	0.79

4. Kết quả thực nghiệm

❖ **Content models + number models**



Classification model	Accuracy	F1-score
Combine Naive Bayes	0.87	0.87
Combine SVM	0.89	0.89
Combine Logistic Regression	0.90	0.90

Q&A

Thank you!