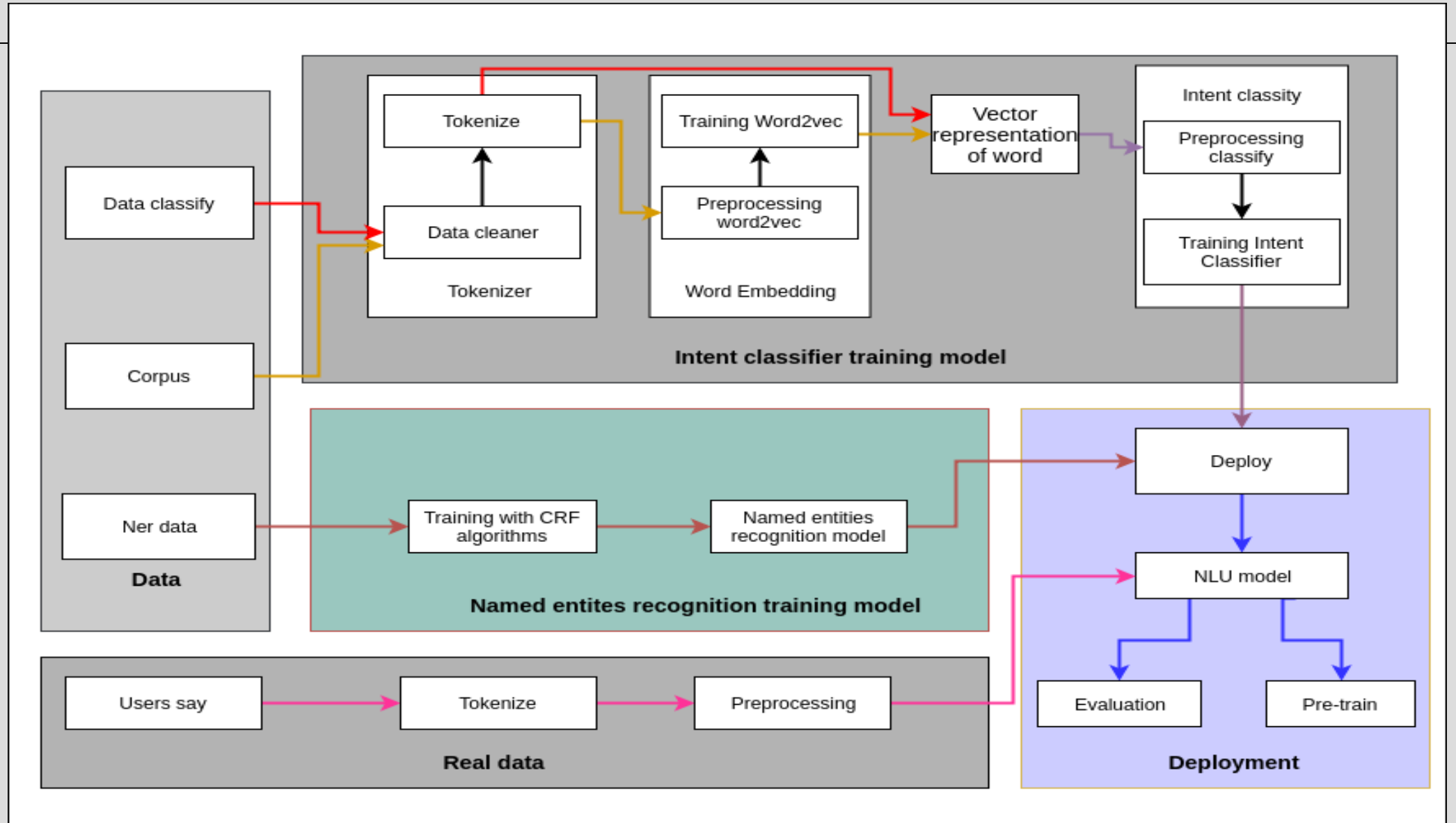
NATURAL LANGUAGE UNDERSTANDING

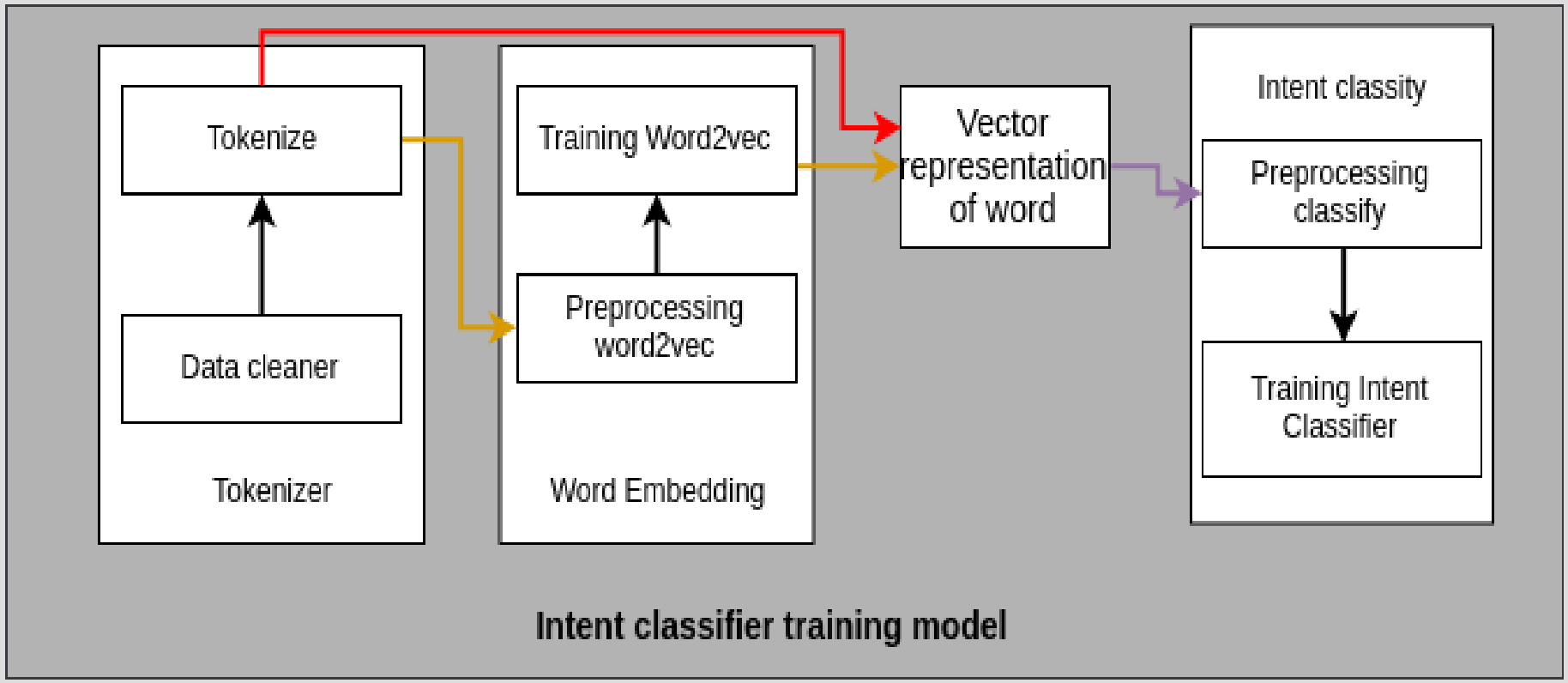
OUTLINES

- Overall architecture
- Intent classifier
- Named entities recognition
- Deployment
- Conclution and future work

OVERALL ARCHITECTURE



INTENT CLASSIFIER

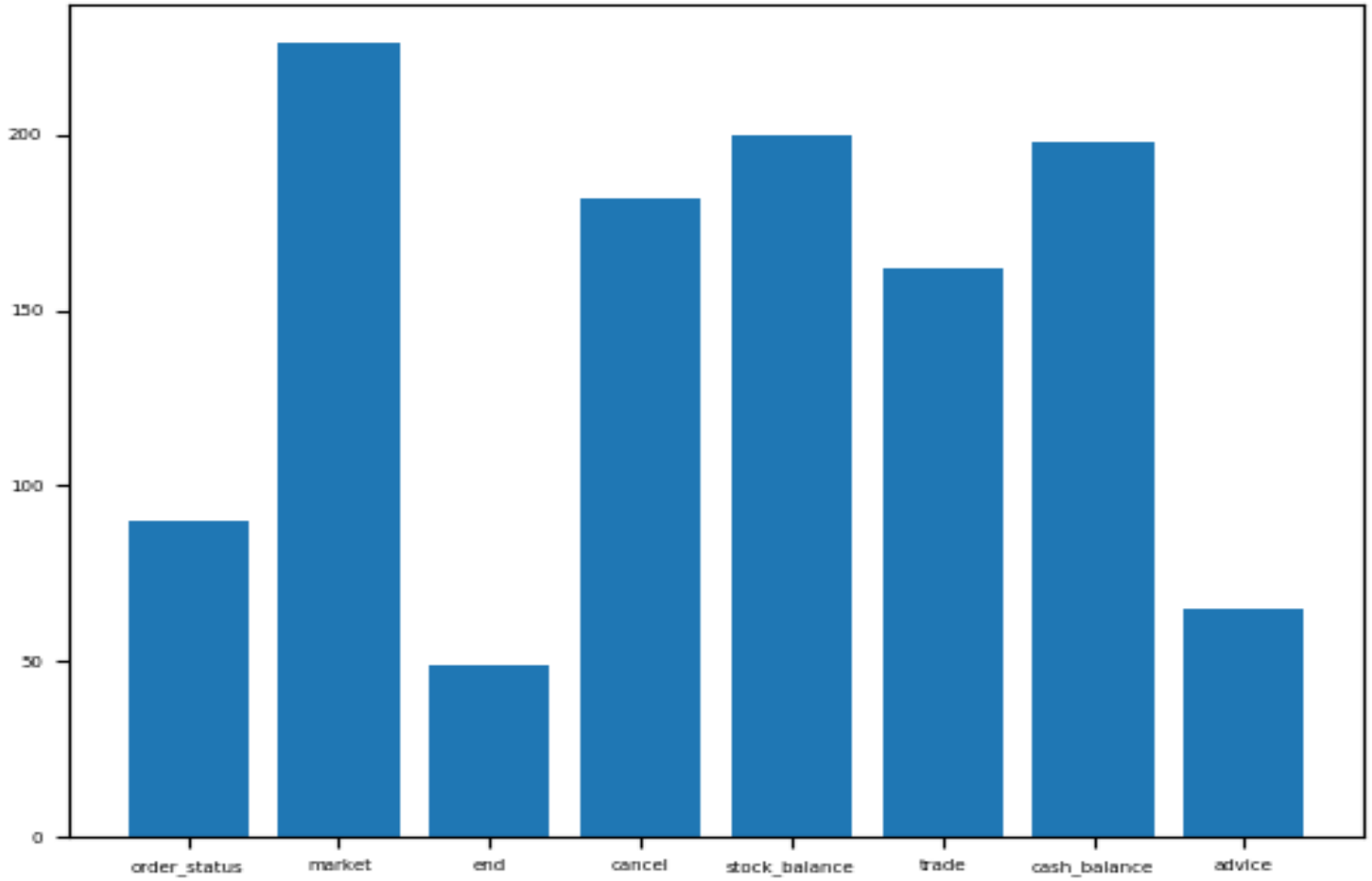


CORPUS

Thị trường có một phiên giao dịch trong biên độ hẹp với sự phân hóa mạnh và thanh khoản duy trì tích cực nhưng những tín hiệu giao dịch ở cuối phiên cho thấy đà tăng đang chững lại và có dấu hiệu quay đầu.

VNIndex đóng cửa vẫn tăng nhẹ 0.63 điểm trong khi hầu hết các chỉ số khác như VN30, HNXIndex giảm điểm cho thấy sự biến động trái chiều và thanh khoản toàn thị trường vẫn duy trì tốt khi đạt hơn 6.000 tỷ đồng trong đó có giao dịch đột biến HNG với qui mô 580 tỷ đồng, VIC 236 tỷ đồng.

TEXT CLASSIFY DATA



trade, đầu tư thêm khối lượng cổ phiếu abc giá

market,thị trường chứng khoán hôm nay thế nào nhỉ?

TOKENIZER

Goal: Clean and tokenize data

Components:

DataCleaner: Remove special character

Remove stopwords

Replace acronym

Tokenizer: Use pyvi library to tokenize Vietnamese words

=> Vocabulary: 2444 words.

WORD EMBEDDING

Goal: Training a model that can build word representation

Components:

PreprocessingWord2vec:

- Create one hot vector for each word
- Create samples word2vec with other window size

Traning word2vec model:

- Using skip-gram method
- Build ANN for training
- Experiment with difference hyperparameters: batch_size, EMBEDDING DIM

INTENT CLASSIFY

Goal: Training a Classifier

Components:

PreprocessingClassifier:

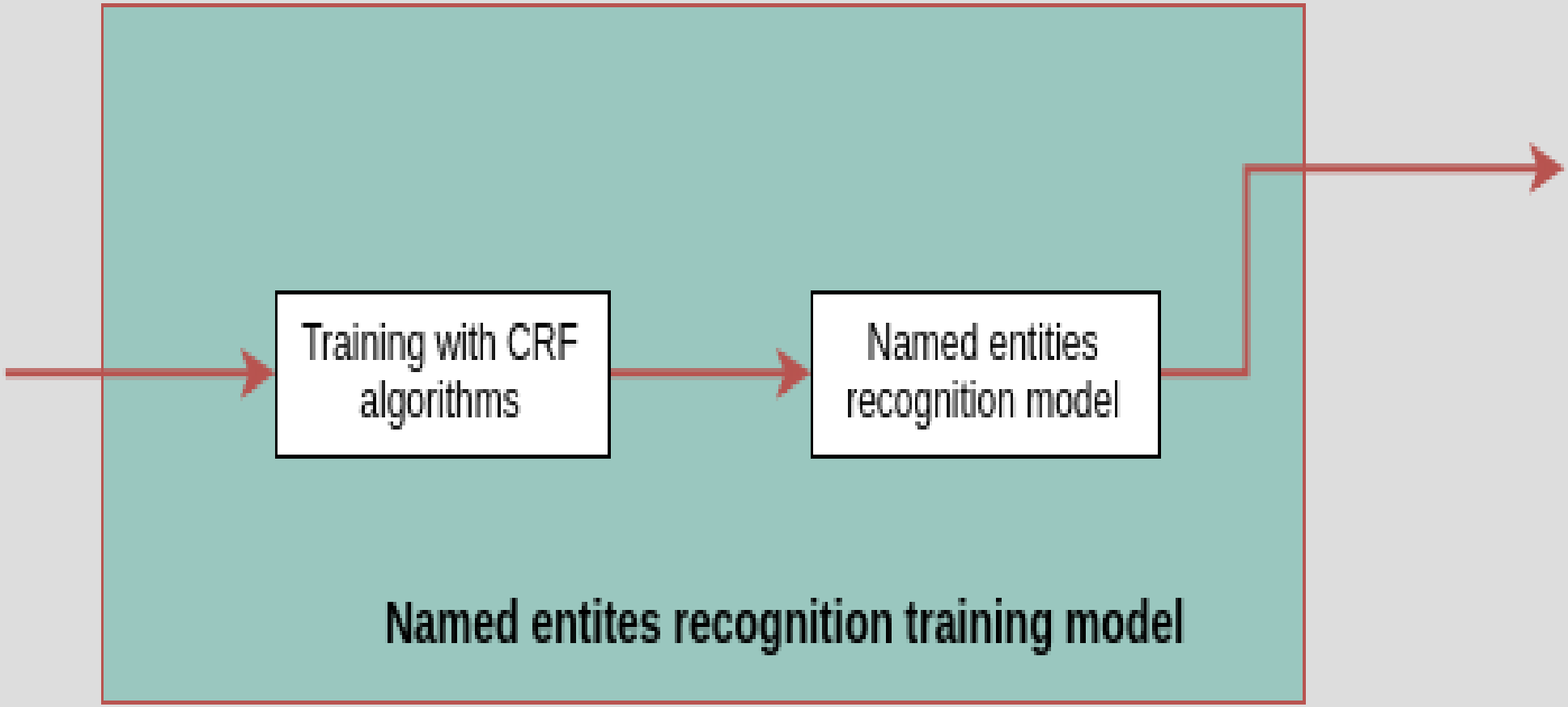
- Use word2vec models
- Tokenize training sentences
- 80% data for training and 20% for testing

Training Classifier:

- Build ANN for training
- Experiment with difference hyperparameters: batch_size, num_units, num_layers

=> Accuracy on test data: 0.9489

NAMED ENTITIES RECOGNITION



DATA

mình P O

muốn V O

chuyển_nhượng V O

khối_lượng N O

741 M quantity

pac Nu symbol

giá N O

9.33 M price

nghìn M O

đồng Nu O

DEPLOYMENT

- Restore Intent classifer model and ner model
- Experiment with user request
- Pre-train models

CONCLUTION AND FUTURE WORK

- Do not have mechanism for evaluate named entities recognition model and word2vec model

- Tokenizer is not the best! => need to build a model for tokenizer

- Experiment with other approaches(LSTM, CNN,...)

-

THANKS FOR YOUR ATTENTION!!!