

GROUP 12





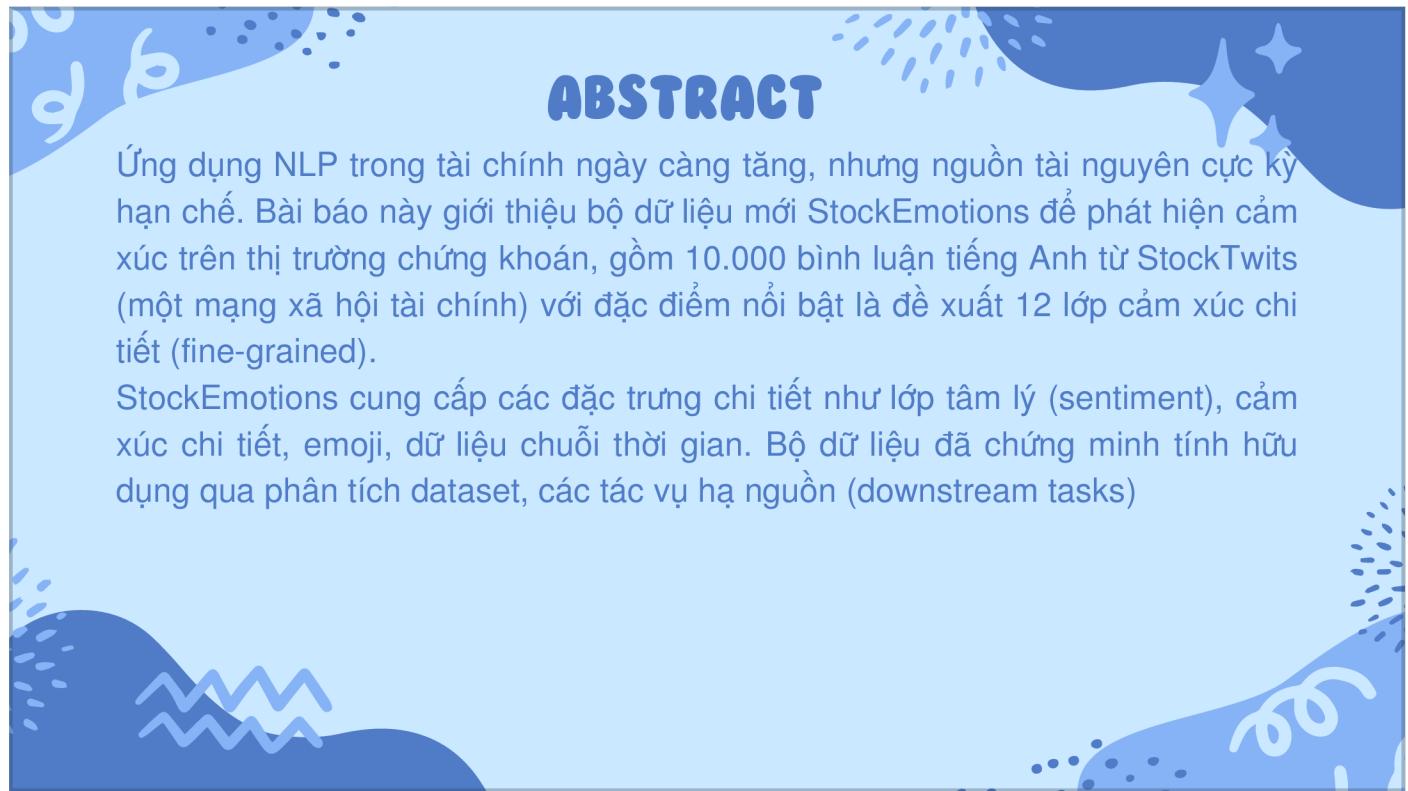
STOCKEMOTIONS

DISCOVER INVESTOR EMOTIONS FOR FINANCIAL SENTIMENT ANALYSIS AND MULTIVARIATE TIME SERIES

ABSTRACT

Ứng dụng NLP trong tài chính ngày càng tăng, nhưng nguồn tài nguyên cực kỳ hạn chế. Bài báo này giới thiệu bộ dữ liệu mới StockEmotions để phát hiện cảm xúc trên thị trường chứng khoán, gồm 10.000 bình luận tiếng Anh từ StockTwits (một mạng xã hội tài chính) với đặc điểm nổi bật là đề xuất 12 lớp cảm xúc chi tiết (fine-grained).

StockEmotions cung cấp các đặc trưng chi tiết như lớp tâm lý (sentiment), cảm xúc chi tiết, emoji, dữ liệu chuỗi thời gian. Bộ dữ liệu đã chứng minh tính hữu dụng qua phân tích dataset, các tác vụ hạ nguồn (downstream tasks)

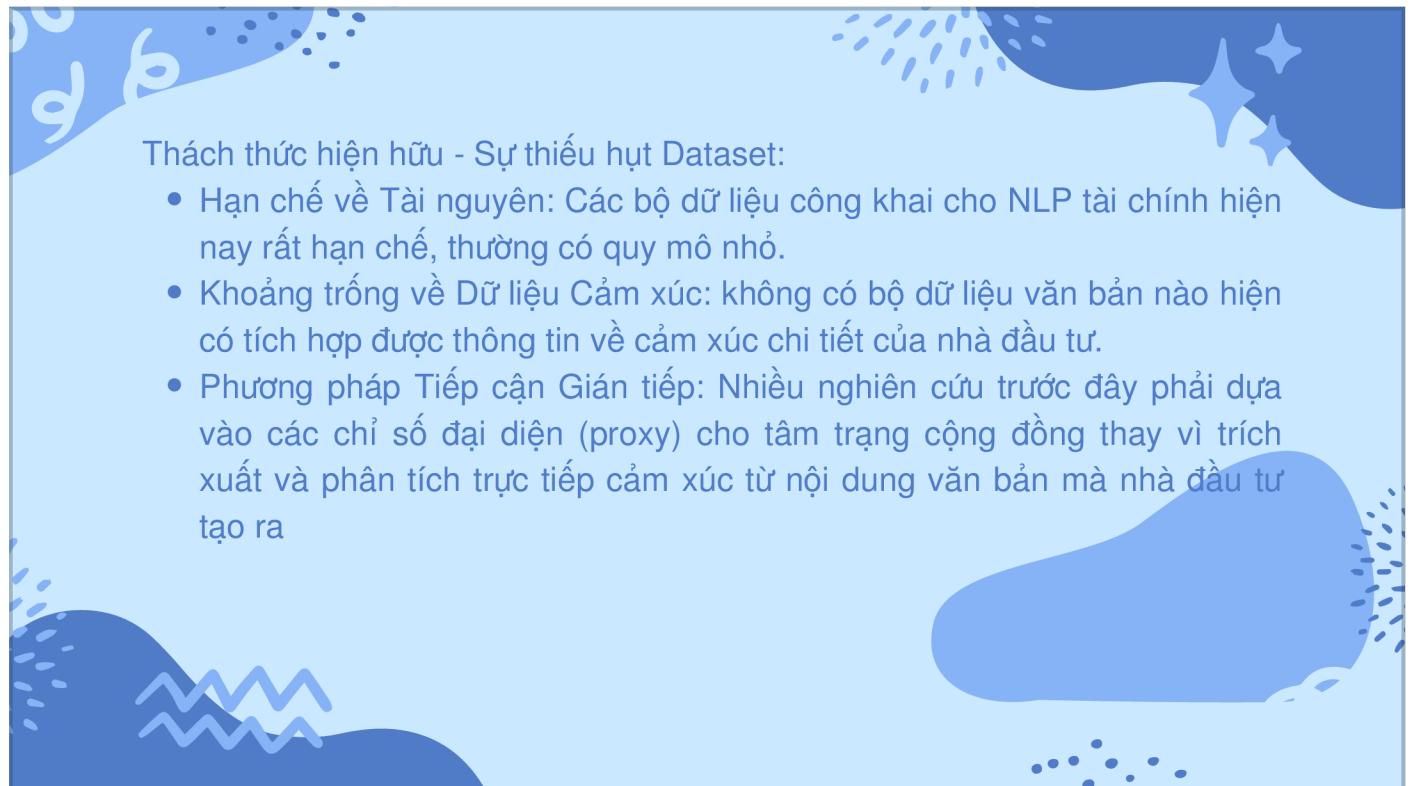


INTRODUCTION

- NLP trong Tài chính là một lĩnh vực nghiên cứu đang phát triển nhanh chóng
- Ứng dụng phổ biến của lĩnh vực này là phân tích tâm lý tài chính từ các nguồn dữ liệu phi cấu trúc như mạng xã hội và tin tức.
- Nền tảng lý thuyết (Tài chính hành vi): Các nghiên cứu về tài chính hành vi nhấn mạnh rằng tâm trạng cộng đồng và tâm lý tài chính đóng vai trò then chốt trong việc hình thành quyết định đầu tư của cá nhân. Các nền tảng mạng xã hội (như StockTwits, Twitter, Reddit) đã trở thành nguồn dữ liệu để phát hiện và theo dõi các tín hiệu tâm lý này trong thời gian thực

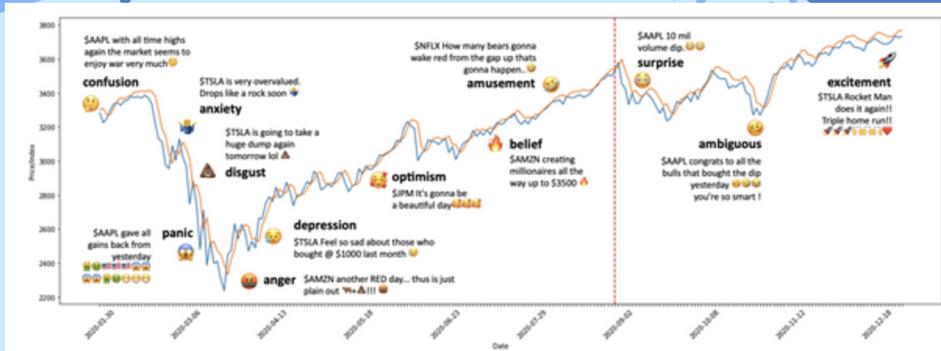
Thách thức hiện hữu - Sự thiếu hụt Dataset:

- Hạn chế về Tài nguyên: Các bộ dữ liệu công khai cho NLP tài chính hiện nay rất hạn chế, thường có quy mô nhỏ.
- Khoảng trống về Dữ liệu Cảm xúc: không có bộ dữ liệu văn bản nào hiện có tích hợp được thông tin về cảm xúc chi tiết của nhà đầu tư.
- Phương pháp Tiếp cận Gián tiếp: Nhiều nghiên cứu trước đây phải dựa vào các chỉ số đại diện (proxy) cho tâm trạng cộng đồng thay vì trích xuất và phân tích trực tiếp cảm xúc từ nội dung văn bản mà nhà đầu tư tạo ra



• Để giải quyết những hạn chế trên, bài báo giới thiệu bộ dữ liệu mới mang tên StockEmotions.

- Quy trình bao gồm việc thu thập 10.000 câu bình luận từ nền tảng StockTwits trong năm 2020 (giai đoạn biến động do COVID-19). Sau đó gán nhãn đa chiều cho 2 lớp tâm lý (Sentiment): Bullish (lạc quan) / Bearish (bi quan) và 12 lớp cảm xúc chi tiết (Fine-grained Emotions): Được gán nhãn thông qua quy trình kết hợp Pre-trained Language Model (PLM) và đánh giá/hiệu chỉnh bởi các chuyên gia tài chính.
- Hệ thống phân loại 12 cảm xúc này được xây dựng cẩn thận, có liên kết với các nghiên cứu tâm lý học hiện có, đảm bảo tính hợp lý và bao quát



Ví dụ Minh họa: Một bình luận như "\$TSLA rocket man going to the moon for sure!!!" không chỉ thể hiện tâm lý 'bullish' mà còn bộc lộ cảm xúc 'excitement' (phản khích). StockEmotions ghi nhận cả hai khía cạnh này, cùng với emoji (🚀🌕) và liên kết đến dữ liệu chuỗi thời gian của TSLA tại thời điểm đó.

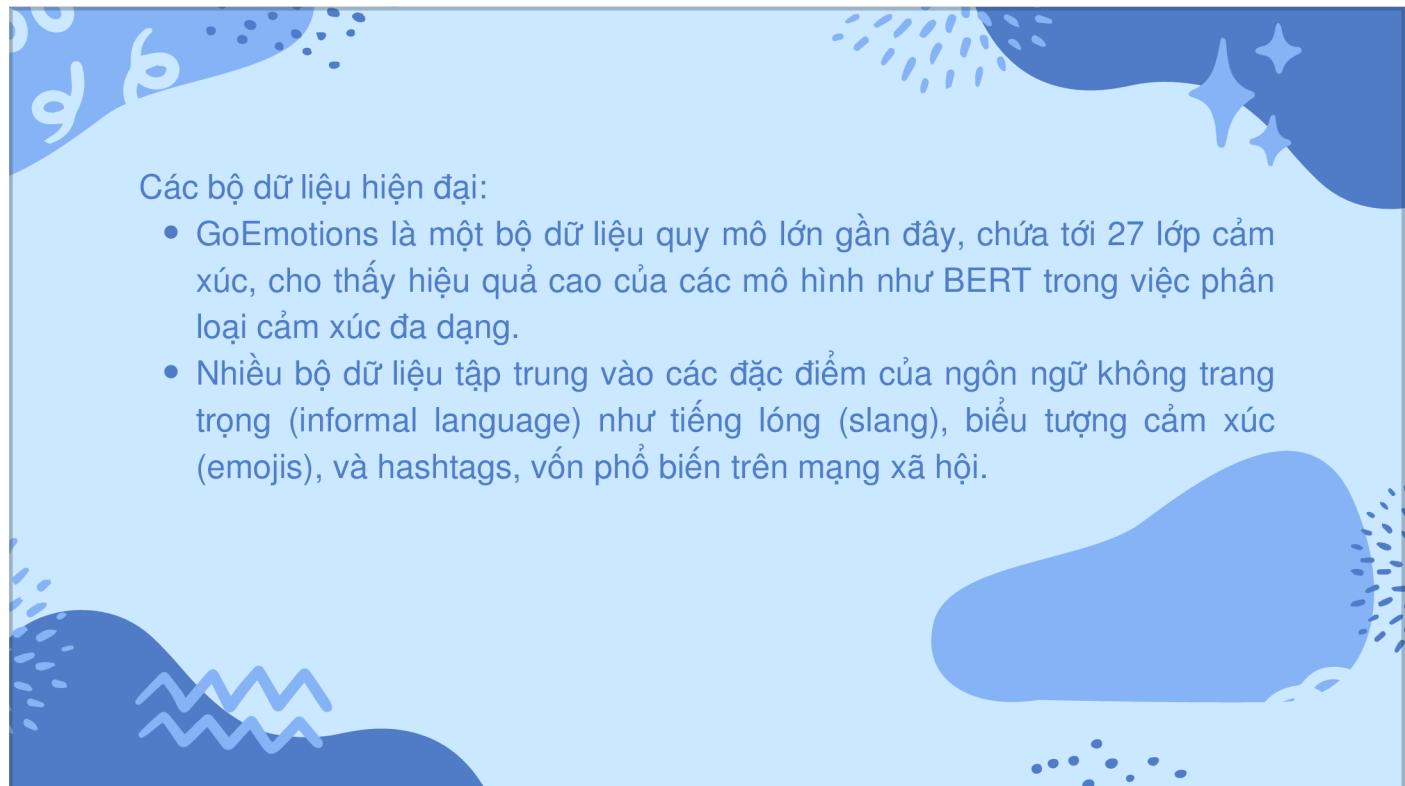
RELATED WORK

Phân loại cảm xúc trong NLP tổng quát

- Nền tảng: Phân loại cảm xúc là một chủ đề cốt lõi và được nghiên cứu sâu rộng trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên (NLP).
- Nguồn Dữ liệu Thường gặp: Dữ liệu thường được thu thập từ các nền tảng mạng xã hội (Twitter, Facebook), các trang tin tức trực tuyến, và các đoạn hội thoại.
 - Các Hệ thống phân loại cảm xúc kinh điển: Mô hình 6 cảm xúc cơ bản của Ekman (1992): Anger, Disgust, Fear, Happiness, Sadness, Surprise.
 - Bánh xe cảm xúc của Plutchik (1980) với 8 cảm xúc cơ bản và các cấp độ khác nhau: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation.

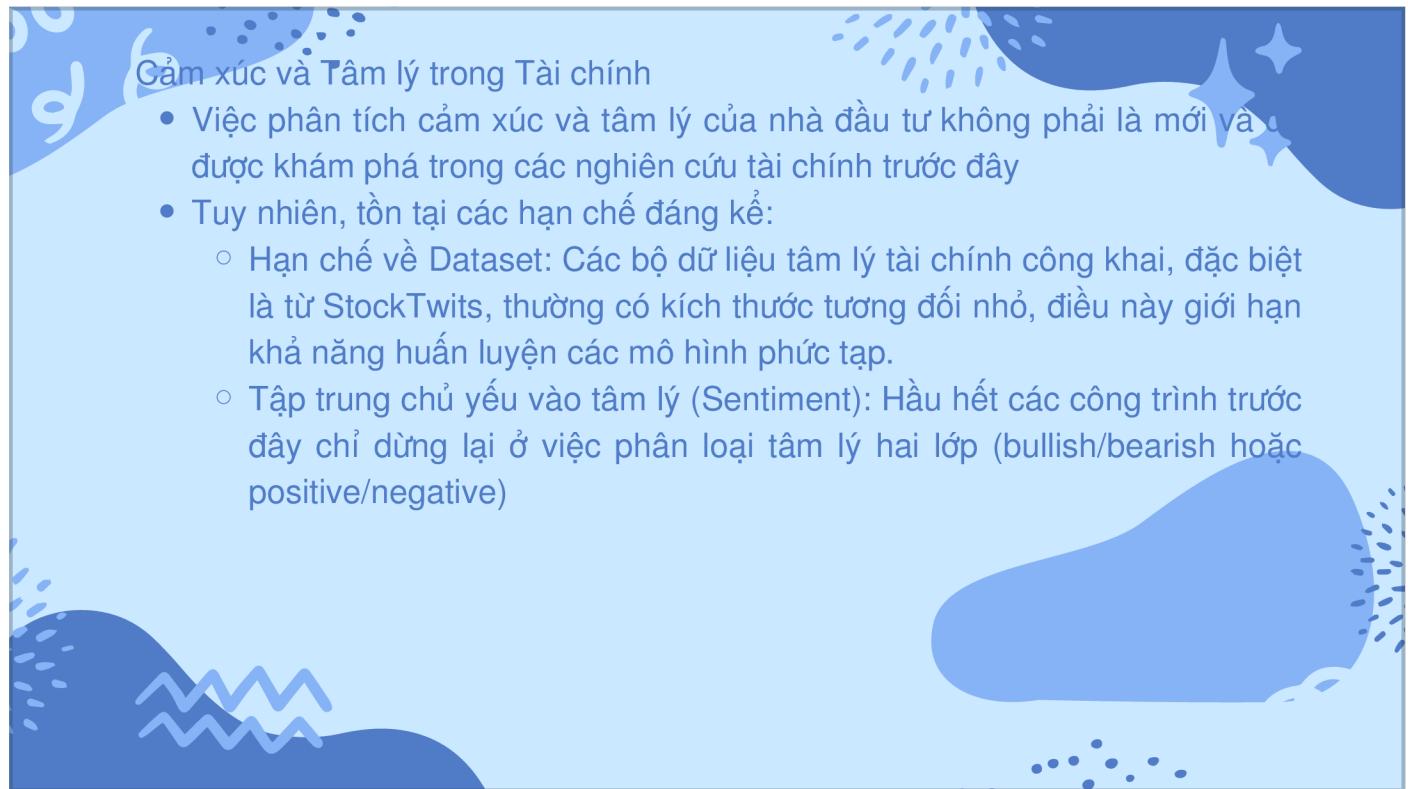
Các bộ dữ liệu hiện đại:

- GoEmotions là một bộ dữ liệu quy mô lớn gần đây, chứa tới 27 lớp cảm xúc, cho thấy hiệu quả cao của các mô hình như BERT trong việc phân loại cảm xúc đa dạng.
- Nhiều bộ dữ liệu tập trung vào các đặc điểm của ngôn ngữ không trọng trọng (informal language) như tiếng lóng (slang), biểu tượng cảm xúc (emojis), và hashtags, vốn phổ biến trên mạng xã hội.



Cảm xúc và Tâm lý trong Tài chính

- Việc phân tích cảm xúc và tâm lý của nhà đầu tư không phải là mới và được khám phá trong các nghiên cứu tài chính trước đây
- Tuy nhiên, tồn tại các hạn chế đáng kể:
 - Hạn chế về Dataset: Các bộ dữ liệu tâm lý tài chính công khai, đặc biệt là từ StockTwits, thường có kích thước tương đối nhỏ, điều này giới hạn khả năng huấn luyện các mô hình phức tạp.
 - Tập trung chủ yếu vào tâm lý (Sentiment): Hầu hết các công trình trước đây chỉ dừng lại ở việc phân loại tâm lý hai lớp (bullish/bearish hoặc positive/negative)



STOCKEMOTIONSDATASET

1. DATA COLLECTION

2. DATA PROCESSING

3. ANNOTATION



DATA COLLECTION

1. Dữ liệu thu thập từ StockTwits, một nền tảng mạng xã hội tài chính tương tự Twitter, nơi người dùng chia sẻ bình luận về cổ phiếu kèm cashtag và tự gán nhãn sentiment.

2. Quy mô:

- 3 triệu bình luận, bao phủ hơn 80% công ty trong S&P 500 theo vốn hóa thị trường.
- Khoảng thời gian: 01/2020 - 12/2020



DATA PROCESSING

1. TOKENIZATION & LENGTH FILTERING

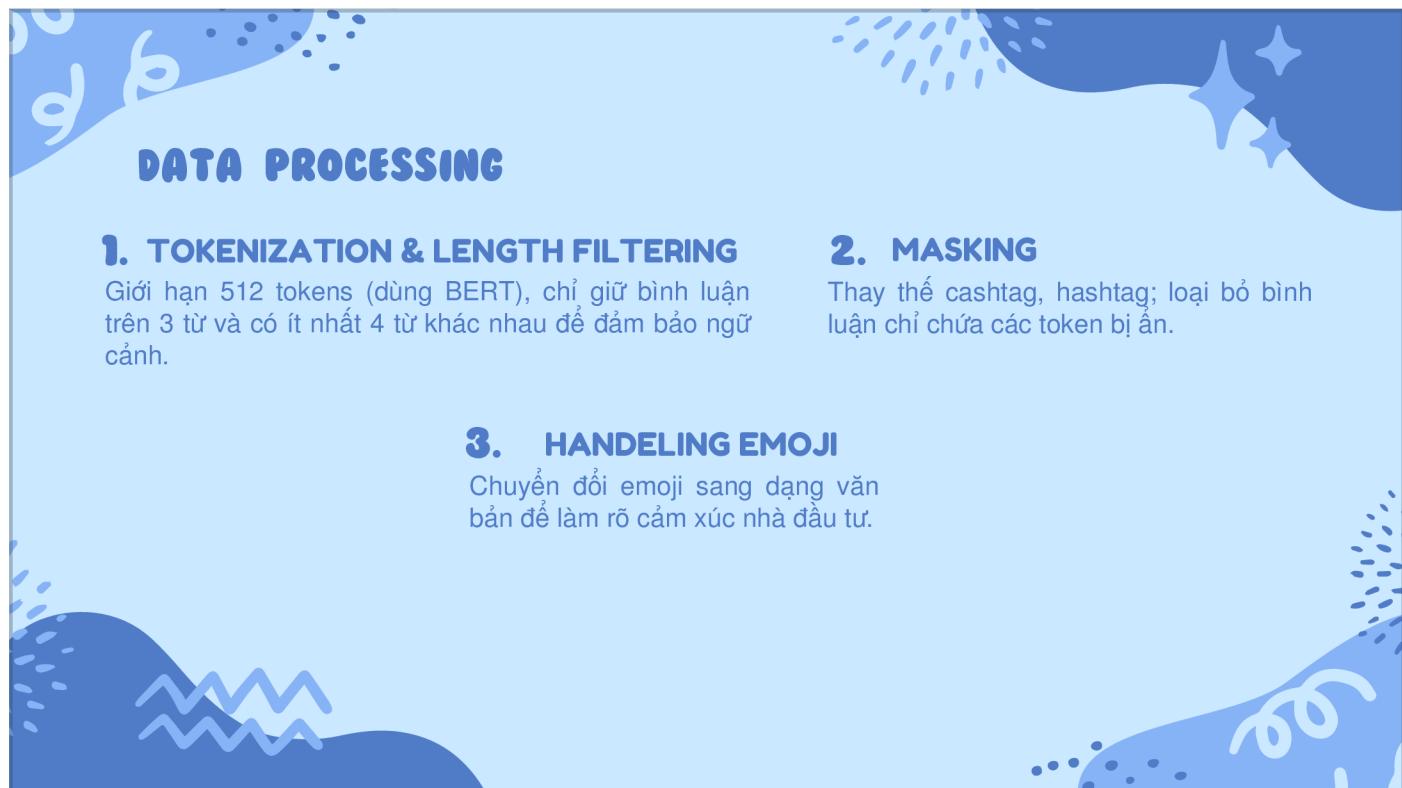
Giới hạn 512 tokens (dùng BERT), chỉ giữ bình luận trên 3 từ và có ít nhất 4 từ khác nhau để đảm bảo ngữ cảnh.

2. MASKING

Thay thế cashtag, hashtag; loại bỏ bình luận chỉ chứa các token bị ẩn.

3. HANDLING EMOJI

Chuyển đổi emoji sang dạng văn bản để làm rõ cảm xúc nhà đầu tư.



ANNOTATION

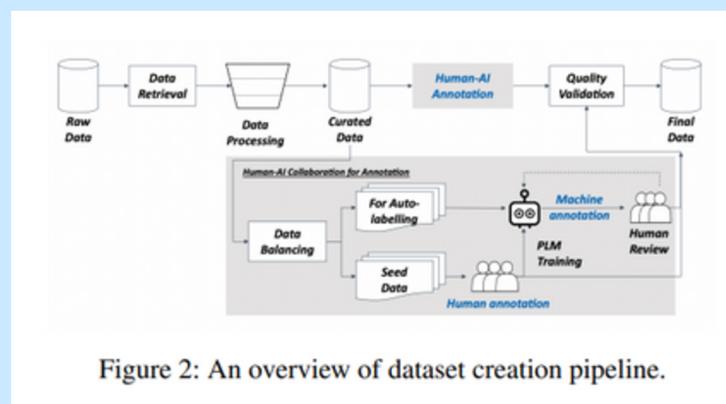
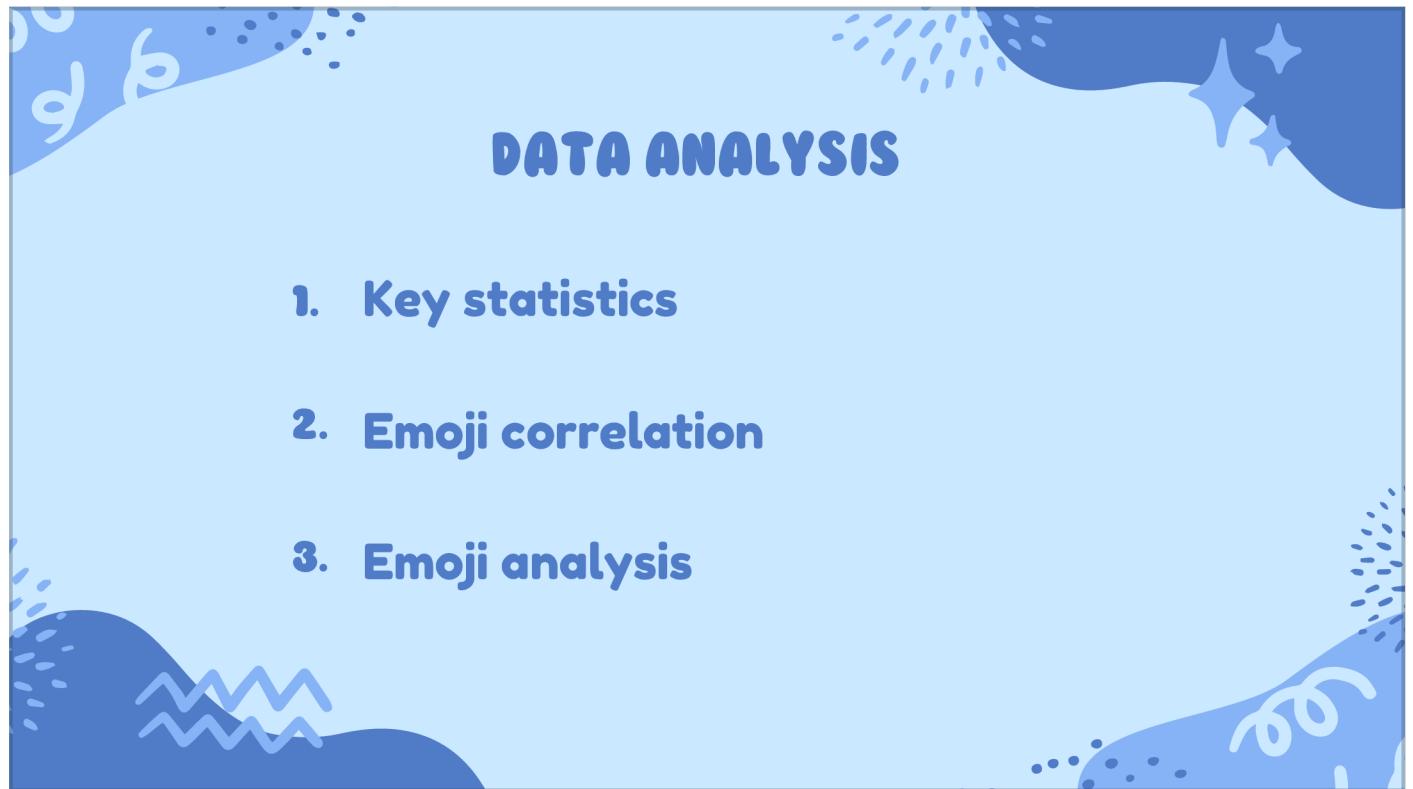


Figure 2: An overview of dataset creation pipeline.

Kết hợp giữa mô hình ngôn ngữ tiền huấn luyện (PLM) và chuyên gia tài chính để gán nhãn cảm xúc.

DATA ANALYSIS

- 1. Key statistics**
- 2. Emoji correlation**
- 3. Emoji analysis**



KEY STATISTIC

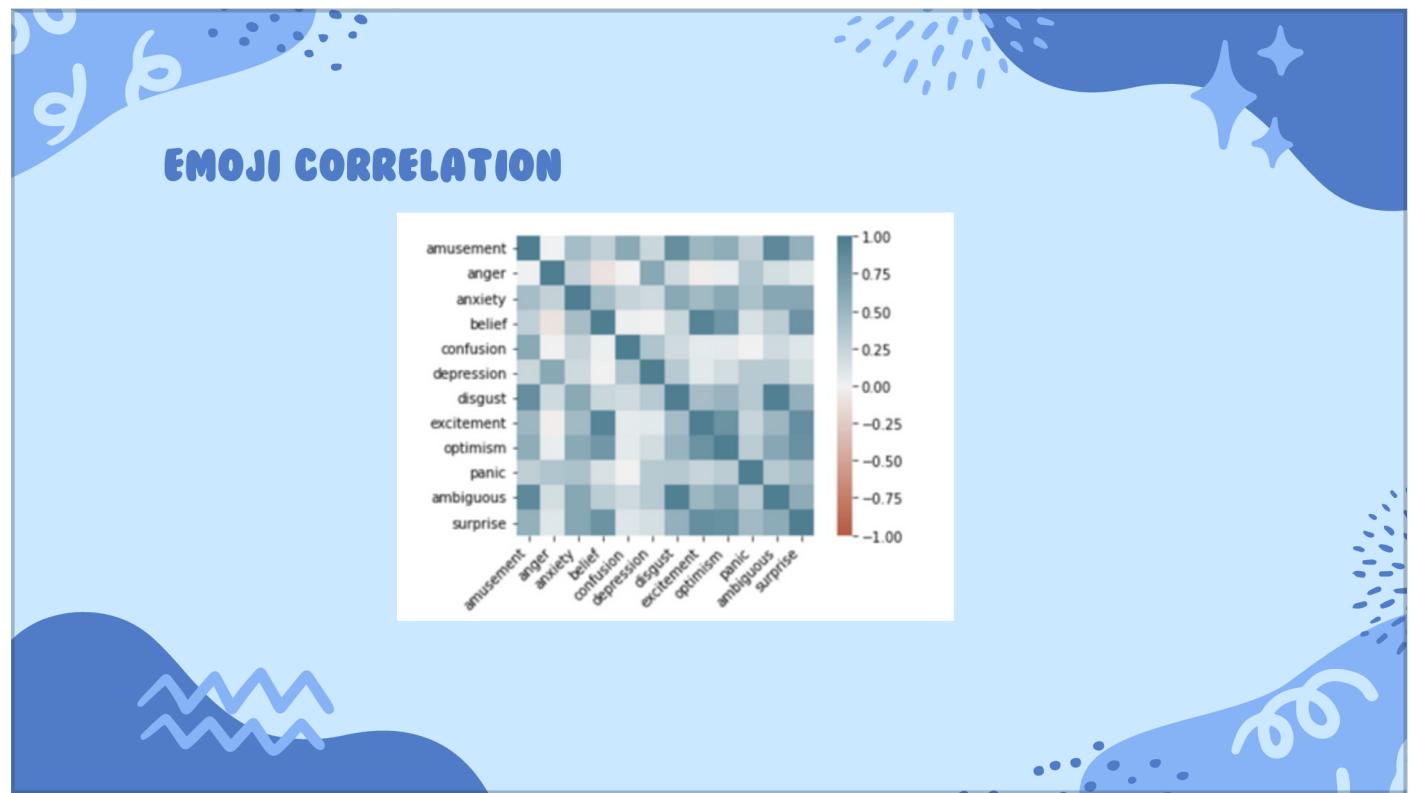
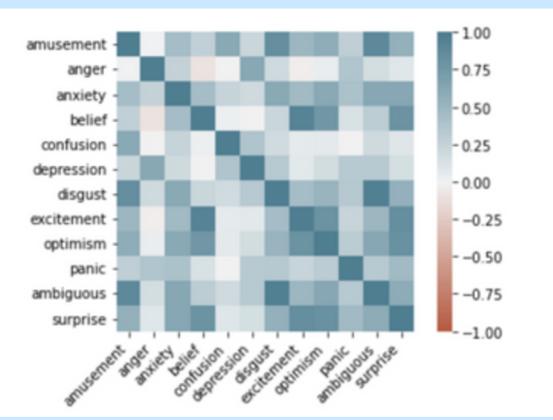
Number of Utterance	10,000
Number of Sentiment	2 - bullish (55%), bearish (45%)
Number of Emotion	12 - ambiguous(9%), amusement(8%), anger(4%), anxiety(14%), belief(9%), confusion(6%), depression(2%), disgust(13%), excitement(14%), optimism (16%), panic(3%), surprise(3%)
Avg. Length	19.2 tokens per utterance
Unique Emoji	761
Time Period	01 Jan 2020 - 31 Dec 2020

KEY STATISTIC

Emotion	Definition	Synonyms	Emoji
amusement	the pleasure that you get from being entertained or from doing something interesting.	enjoyment, delight, laughter, pleasure, fun	😂
anger	a strong feeling of being upset or annoyed because of something wrong, unfair, cruel, or unacceptable.	rage, outrage, fury, wrath, irritation	😡
anxiety	a feeling of nervousness or worry about what might happen	nervousness, alarm, worry, tension, uneasiness	😨
belief	a feeling of certainty that something exists, is true, or is good, associated with the company's operation.	trust, faith, confidence, conviction, reliance	🔥
confusion	a refusal or reluctance to believe	scepticism, doubt, disbelief, distrust, uncertainty	🤔
depression	a state of feeling sad, extreme gloom, inadequacy, and inability to concentrate	sadness, despair, giving up, hopelessness, gloom	😢
disgust	a feeling of very strong dislike or disapproval.	loathing, dislike, hatred, sicken, abomination	🤮
excitement	a feeling of having great enthusiasm, strong belief, intense enjoyment, or great eagerness.	enthusiasm, passion, cheerfulness, heat	🚀
optimism	a feeling of being hopeful about the future or about the success of something in particular.	hope, wish, desire, want, positiveness	🏆
panic	a very strong feeling of anxiety or fear, which makes you act without thinking carefully.	horror, terror, fear, dismay, terrify	😱
surprise	a feeling caused by something that is unexpected or unusual. (e.g. earning surprise)	amazement, astonishment, shock, revelation	😲
ambiguous	unclassified emotions in the list or when the target of emotion is confused.	(subject to annotator's understanding of the text)	🤔

Table 4: Emotion Definition given to the annotators.

EMOJI CORRELATION



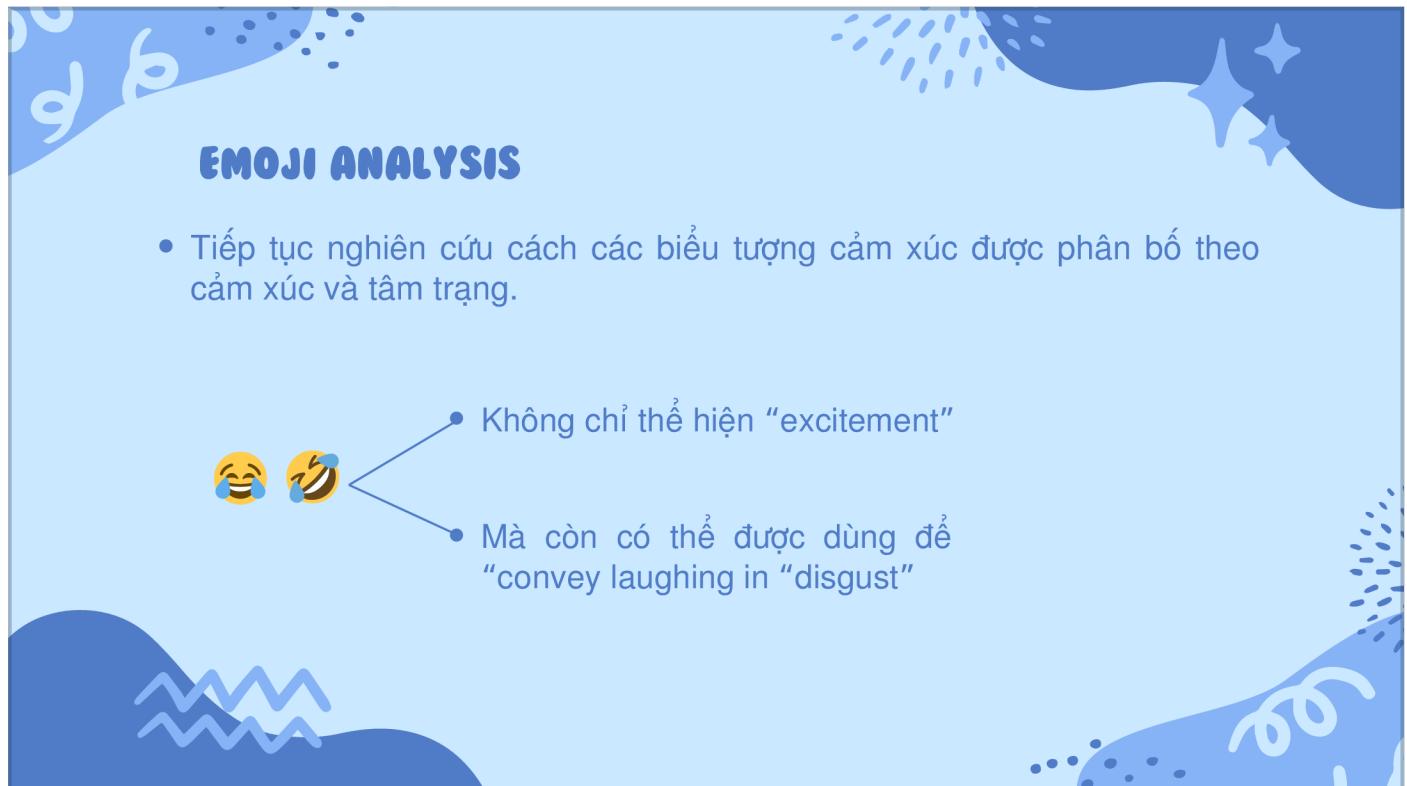
EMOJI ANALYSIS

- Tiếp tục nghiên cứu cách các biểu tượng cảm xúc được phân bố theo cảm xúc và tâm trạng.



• Không chỉ thể hiện “excitement”

• Mà còn có thể được dùng để
“convey laughing in “disgust”





MODELLING AND EXPERIMENTS

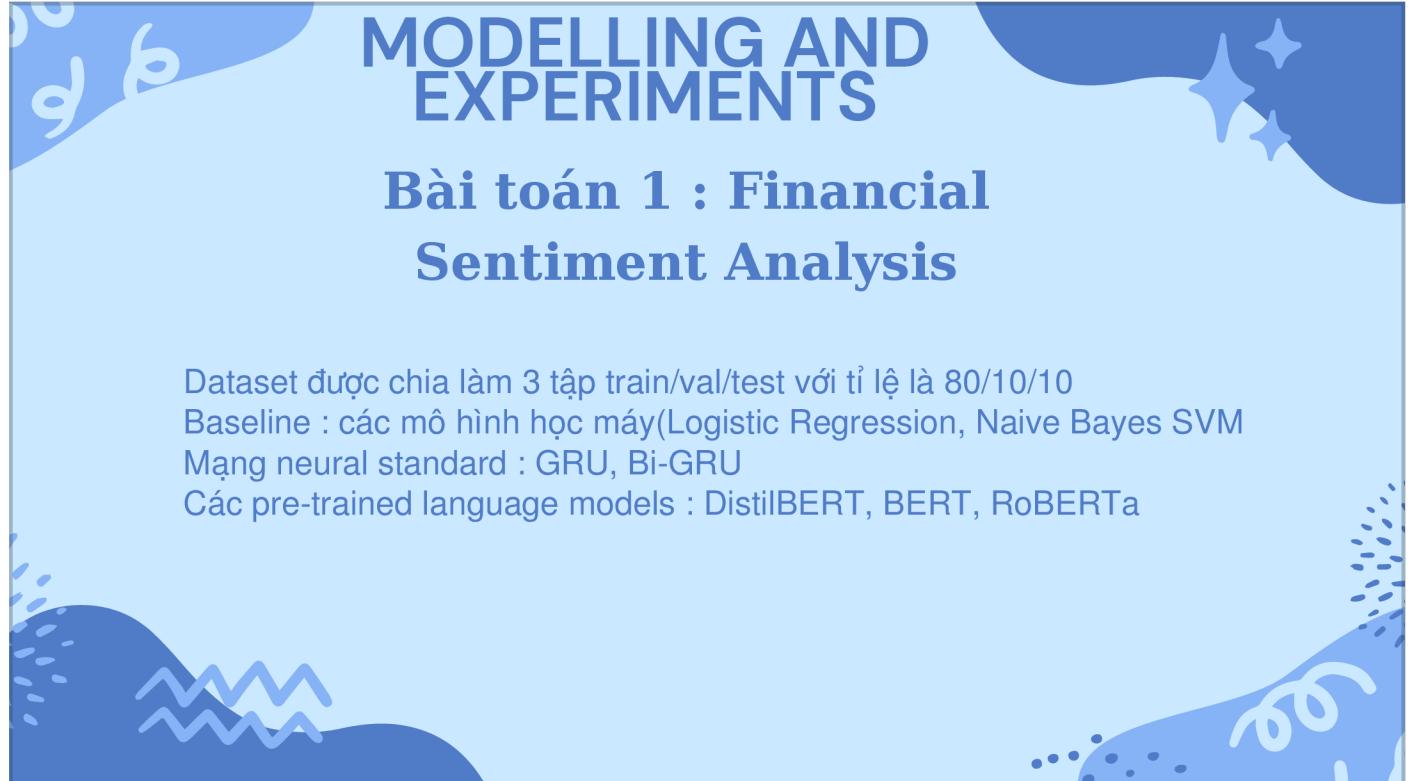
Bài toán 1 : Financial Sentiment Analysis

Dataset được chia làm 3 tập train/val/test với tỉ lệ là 80/10/10

Baseline : các mô hình học máy(Logistic Regression, Naive Bayes SVM

Mạng neural standard : GRU, Bi-GRU

Các pre-trained language models : DistilBERT, BERT, RoBERTa



RESULT

Bài toán 1 : Financial Sentiment Analysis

Model (F1-score)	Sentiment			Emotion												
	bear.	bull.	avg.	ambg.	amus.	angr.	anxt.	belf.	cnfs.	dprs.	disg.	exct.	optim.	panc.	surp.	avg.
LogitReg.	0.71	0.77	0.74	0.12	0.29	0.44	0.37	0.29	0.48	0.24	0.29	0.39	0.31	0.24	0.15	0.32
NBSVM.	0.71	0.78	0.75	0.10	0.27	0.45	0.30	0.36	0.46	0.21	0.34	0.42	0.29	0.29	0.22	0.33
GRU.	0.72	0.79	0.76	0.20	0.31	0.21	0.41	0.15	0.46	0.19	0.33	0.39	0.38	0.43	0.06	0.34
Bi-GRU.	0.73	0.78	0.76	0.22	0.33	0.49	0.39	0.30	0.54	0.29	0.39	0.43	0.32	0.41	0.06	0.36
DistilBERT.	0.79	0.83	0.81	0.12	0.37	0.56	0.42	0.42	0.51	0.29	0.43	0.51	0.42	0.48	0.21	0.42
BERT.	0.79	0.83	0.81	0.27	0.30	0.59	0.46	0.37	0.50	0.22	0.37	0.48	0.40	0.41	0.41	0.40
RoBERTa.	0.78	0.82	0.80	0.09	0.25	0.13	0.44	0.29	0.50	0.21	0.43	0.44	0.39	0.11	0.21	0.39

Table 2: F1-score_{micro} results on the test set for classification task. Emotion labels are in alphabetical order; *ambiguous(ambg.)*, *amusement(amus.)*, *anger(angr.)*, *anxiety(anxt.)*, *belief(belf.)*, *confusion(cnfs.)*, *depression(dprs.)*, *disgust(disg.)*, *excitement(exct.)*, *optimism(optim.)*, *panic(panc.)*, and *surprise(surp.)*.

MODELLING AND EXPERIMENTS

Bài toán 2 : Dự đoán price index theo thời gian thực đa biến

Dữ liệu được lấy từ API của Yahoo Finance chia dữ liệu thành:

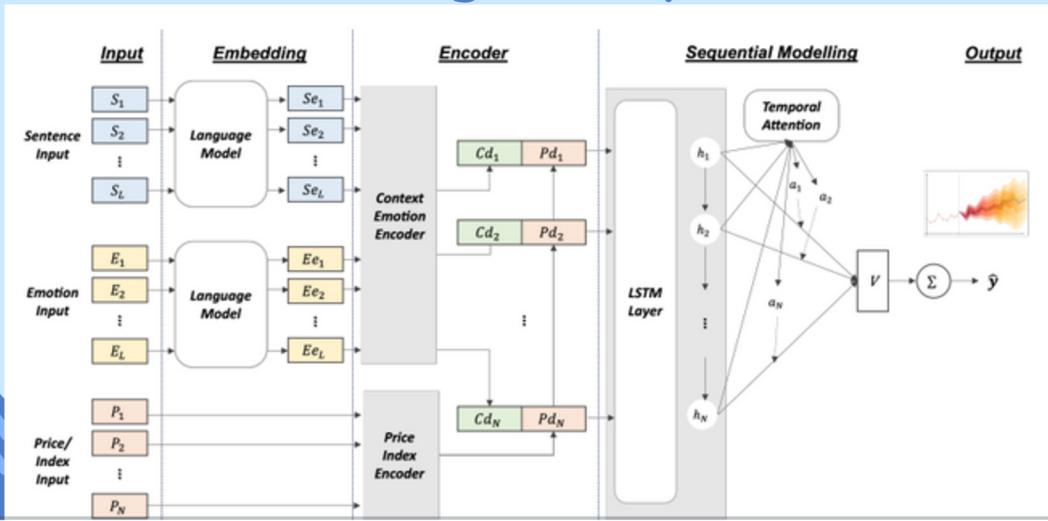
- Train data : (1/1/2020-3/9/2020) 67%
- Test data : (4/9/2020-31/12/2020) 33%

Dữ liệu bao gồm : sentence, emotion, price index

- Sentence embedding bằng BERT
- Emotion embedding bằng GloVe
- Kết hợp với price index để tạo ra bộ encoder

MODELLING AND EXPERIMENTS

Bài toán 2 : Dự đoán price index theo thời gian thực đa biến



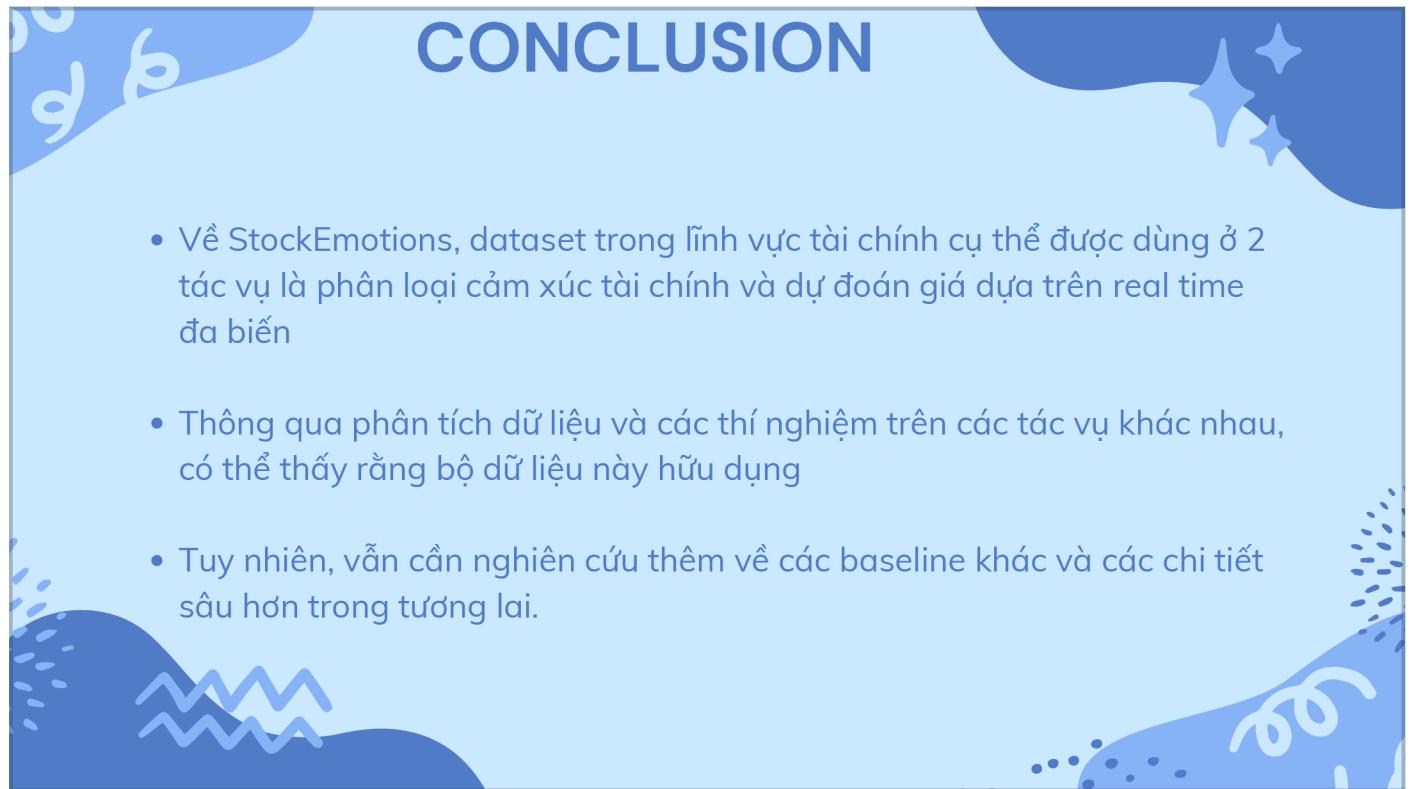
RESULT

Bài toán 2 : Dự đoán price index theo thời gian thực đa biến

Model	window size = 3			window size = 5		
	25	50	100	25	50	100
only index	1.13	1.53	1.83	1.15	2.07	1.99
+ text	2.18	2.30	1.49	0.89	1.32	1.53
+ text + emo.	1.06	1.00	1.39	<u>0.83</u>	1.08	1.48

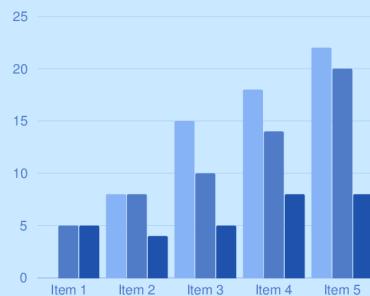
CONCLUSION

- Về StockEmotions, dataset trong lĩnh vực tài chính cụ thể được dùng ở 2 tác vụ là phân loại cảm xúc tài chính và dự đoán giá dựa trên real time đa biến
- Thông qua phân tích dữ liệu và các thí nghiệm trên các tác vụ khác nhau, có thể thấy rằng bộ dữ liệu này hữu dụng
- Tuy nhiên, vẫn cần nghiên cứu thêm về các baseline khác và các chi tiết sâu hơn trong tương lai.



RESULT

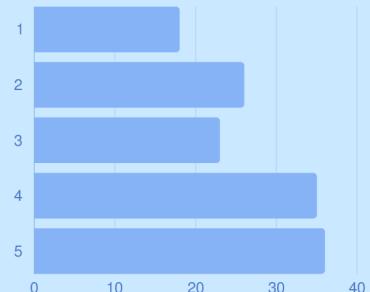
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc a ultricies tortor. In vestibulum vitae velit nec viverra. Proin non ultrices ex. Integer mattis dui vel pretium euismod. Morbi dictum diam nec massa porttitor aliquet.





CHART

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc a ultricies tortor. In vestibulum vitae velit nec viverra. Proin non ultrices ex. Integer mattis dui vel pretium euismod. Morbi dictum diam nec massa porttitor aliquet.



QUESTION TIME





THANK YOU