UNIVERSITÀ DEGLI STUDI DELL'AQUILA

Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica

Corso di Laurea Magistrale in Data Science Applicata

Development of a Scalable Multi-Agent System for Financial Sentiment Analysis Using RabbitMQ

Olusegun Emmanuel Ajibola

RELATORE: Prof. DE GASPERIS GIOVANNI

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

INTRODUCTION

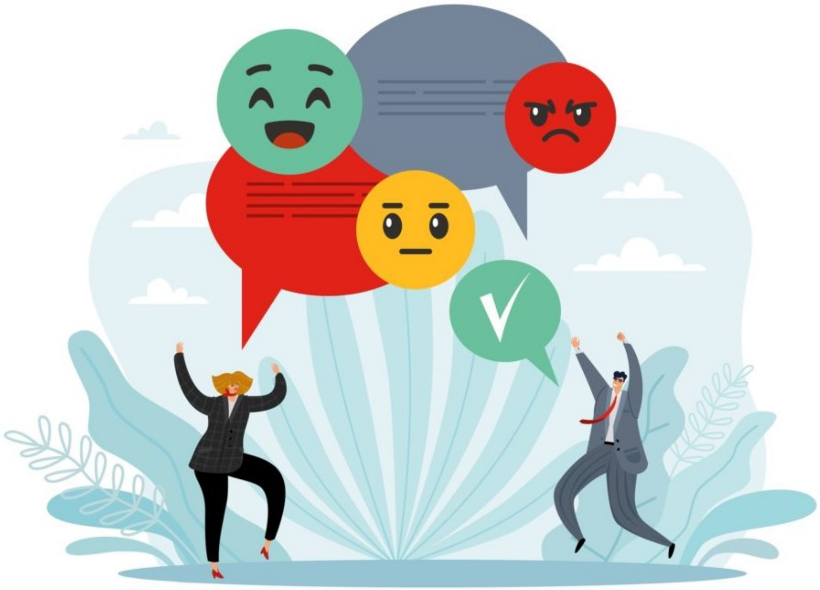


TABLE OF CONTENTS

OVERVIEW

Building a Scalable Multi-Agent System (MAS) for Sentiment Analysis Using RabbitMQ

- Tools

PyCharm
groq
jupyter
PyTorch
NVIDIA
CUDA
git
GitHub
kaggle
RabbitMQ
scikit learn
IOTA

OBJECTIVE

1) Compare and evaluate the performance of state-of-the-art methodologies on financial news data.

2) Develop a scalable MAS architecture tailored for financial applications.

3) Implement the execution of distributed ledger transactions as response agents in the MAS architecture.

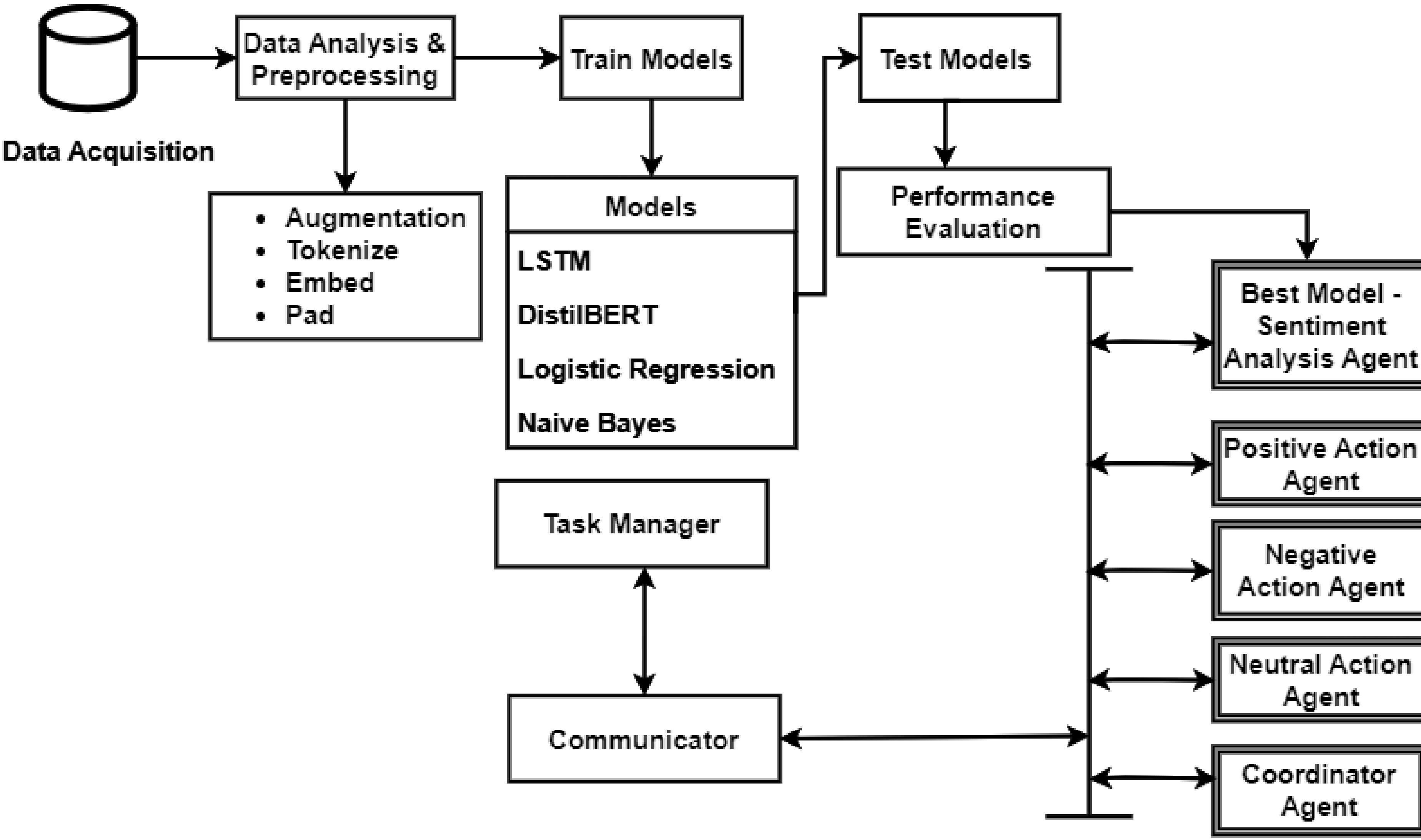
4) Assess the potential benefits and limitations of implementing such systems in the financial industry.

APPLICATIONS

- Financial investment and trading
- social media monitoring
- chatbot enhancement
- customer feedback management e.t.c

METHODOLOGY

Architecture



METHODOLOGY

Data Preprocessing

- Financial PhraseBank from P. Malo et al. (2014) "*Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts*".

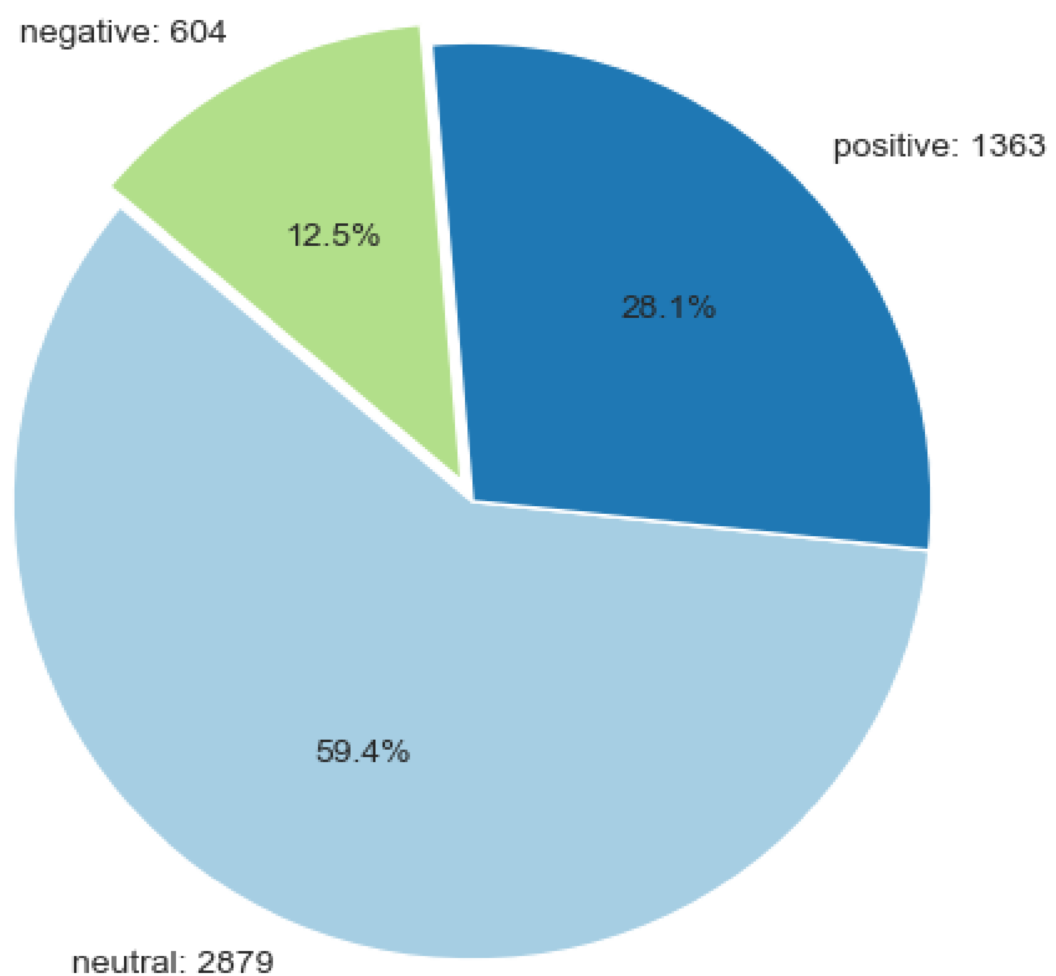
S/No	News	Sentiment
1	The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility ; contrary to earlier layoffs the company contracted the ranks of its office workers , the daily Postimees reported .	negative
2	According to Gran , the company has no plans to move all production to Russia , although that is where the company is growing .	neutral
3	With the new production plant the company would increase its capacity to meet the expected increase in demand and would improve the use of raw materials and therefore increase the production profitability .	positive

TABLE 4.1: Sample news in the dataset.

METHODOLOGY

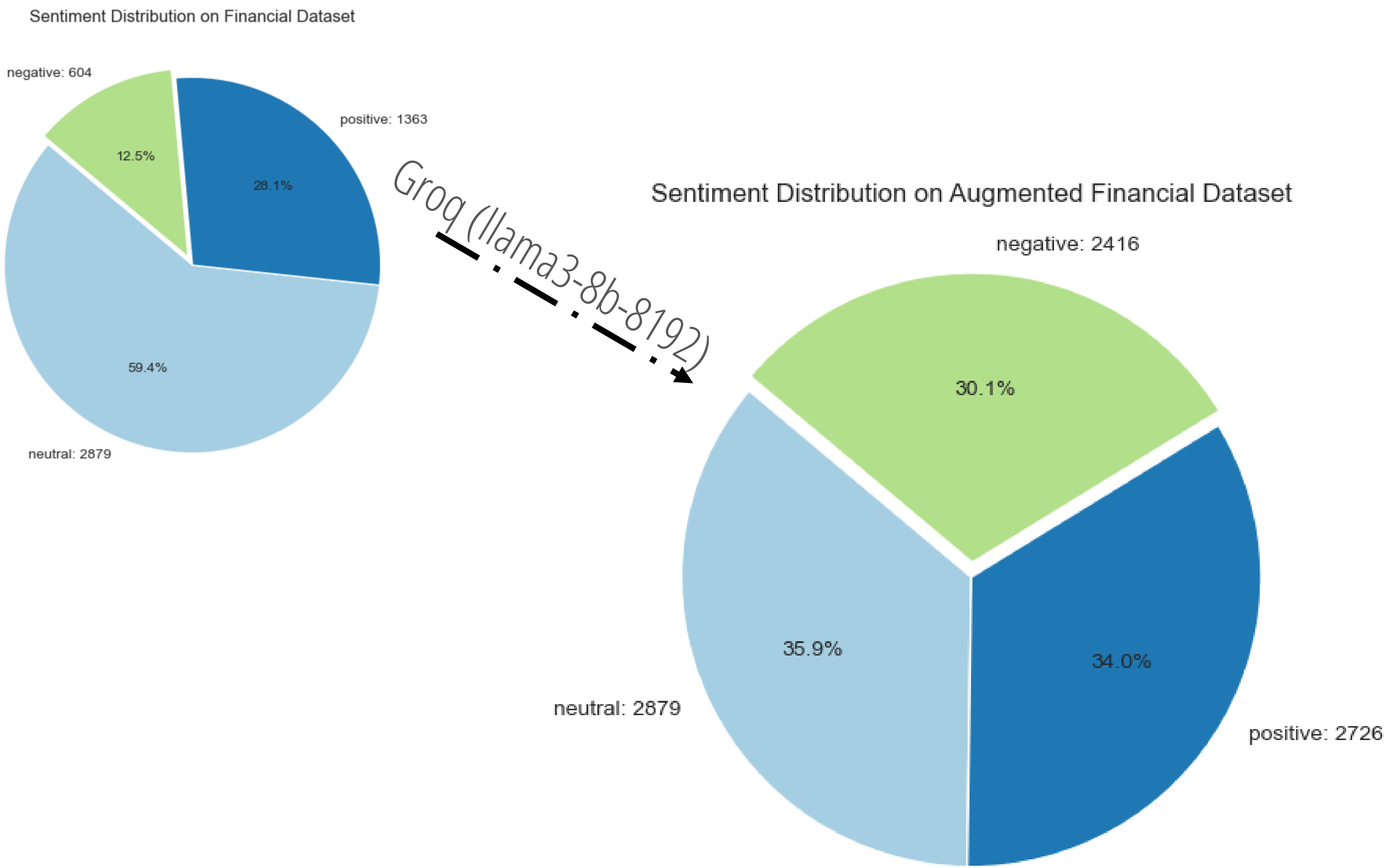
Data Preprocessing

Sentiment Distribution on Financial Dataset



METHODOLOGY

Data Preprocessing: Data Augmentation



Word Count and Encodings

S/No	Label	Encoding
1	positive	2
2	neutral	1
3	negative	0

TABLE 4.3: Label Encoding

Tokenized Word[a]	Embedding (count)
according	1
although	1
company	2
gran	1
growing	1
move	1
plans	1
production	1
russia	1

[a] Original text: According to Gran , the company has no plans to move all production to Russia , although that is where the company is growing .

TABLE 4.5: NB Model Embeddings using Count Vectorizer (word count)

- GloVe

Probability and Ratio	$k = \text{solid}$	$k = \text{gas}$	$k = \text{water}$	$k = \text{fashion}$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

TABLE 3.2: Co-occurrence probabilities for target words ice and steam with selected context words from a 6 billion token corpus. [45]

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f\left(X_{ij}\right) \left(w_i^\top \widetilde{w}_j + b_i + \widetilde{b}_j - \log(X_{ij}) \right)^2 .$$

- BERT and **DistilBERT**

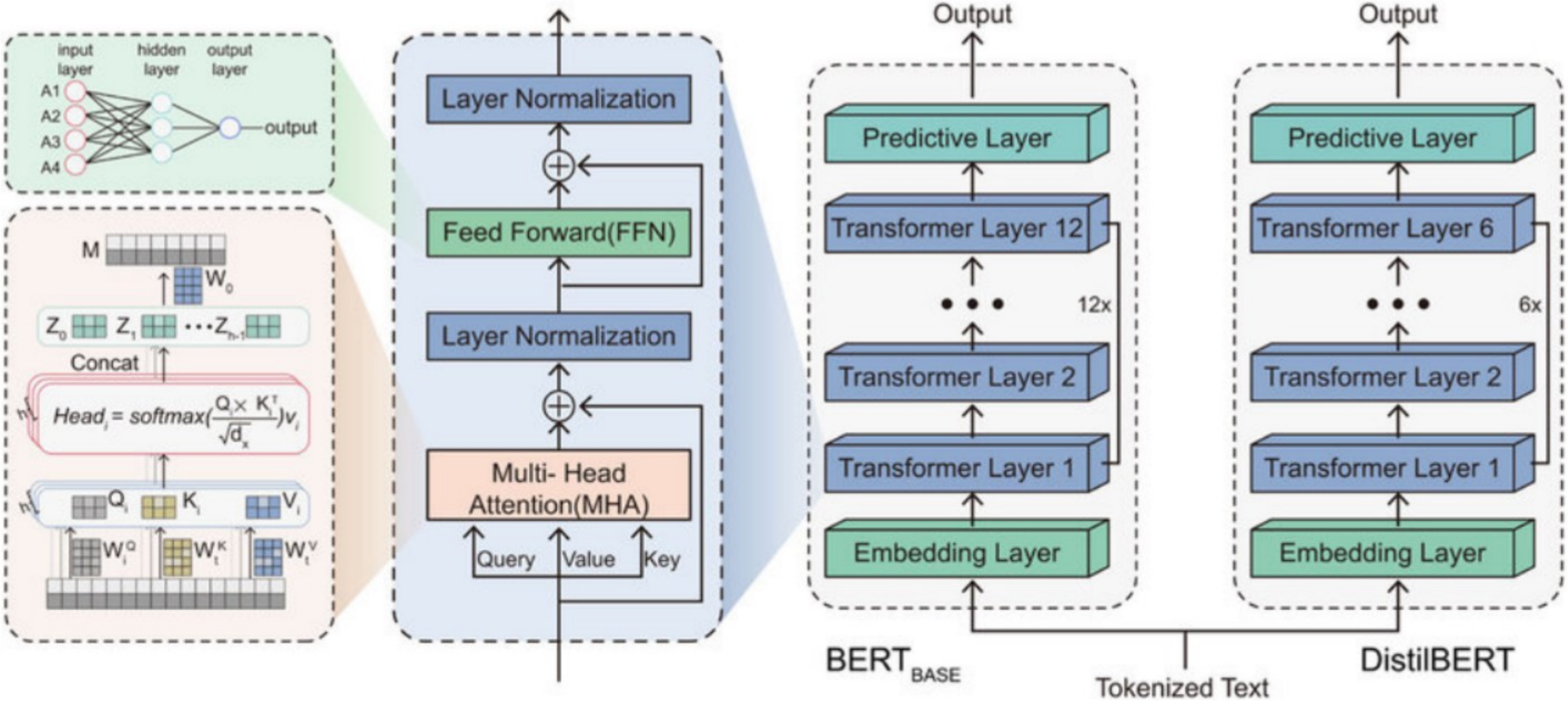


FIGURE 3.4: Diagram of BERT BASE and DistilBERT model architecture [35].

METHODOLOGY

Text Preprocessing: Feature Representations & Embeddings

- BERT and **DistilBERT**

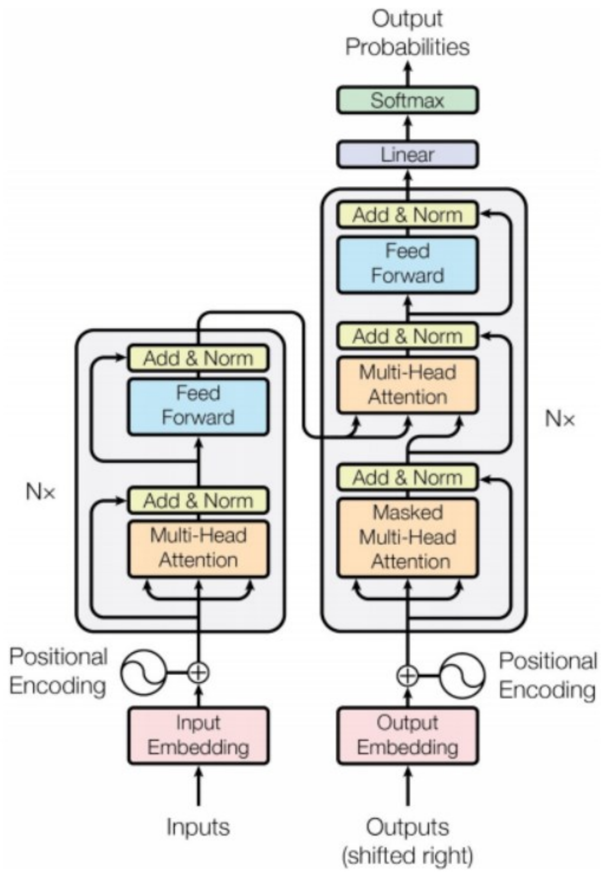


FIGURE 3.7: Model Architecture of a Transformer [57].

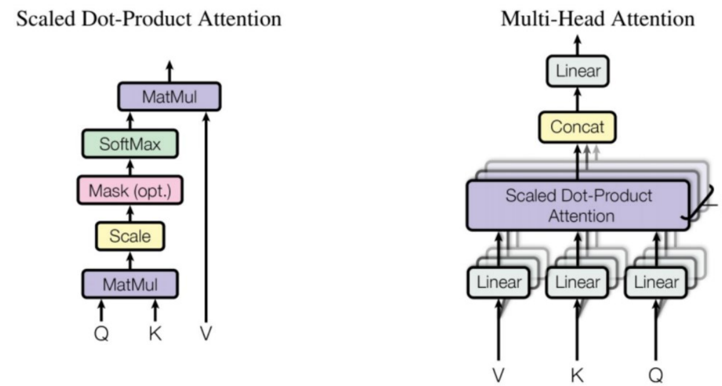


FIGURE 3.8: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel [57].

$$\mathrm{PE}_{(p,2i)} = \sin\left(\frac{p}{10000^{2i/d}}\right), \tag{1}$$

$$\mathrm{PE}_{(p,2i+1)} = \cos\left(\frac{p}{10000^{2i/d}}\right), \tag{2}$$

$$Q = XW^Q, K = XW^K, V = XW^V \tag{3}$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O \tag{5}$$

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{6}$$

METHODOLOGY

Text Preprocessing: Feature Representations & Embeddings

- DistilBERT

S/No	Tokens	Embeddings	S/No	Tokens	Embeddings
0	CLS	101	13	production	2537
1	According	2429	14	to	2000
2	to	2000	15	Russia	3607
3	Gran	12604	16	,	1010
4	,	1010	17	although	2348
5	the	1996	18	that	2008
6	company	2194	19	is	2003
7	has	2038	20	where	2073
8	no	2053	21	the	1996
9	plans	3488	22	company	2194
10	to	2000	23	is	2003
11	move	2693	24	growing	3652
12	all	2035	25	.	1012
			26	SEP	102

TABLE 4.2: Token Embeddings using DistilBERT

- The model is defined as:

$$\Pr(y = i|X) = \frac{e^{w_i^T X + b_i}}{\sum_{j=1}^{k} e^{w_j^T X + b_j}}.$$

- The cross-entropy loss function is employed in the training process of the multinomial logistic regression:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} y_{ij} \log(\Pr(y = j|X_i)).$$

- The model is defined as:

Naive Bayes estimates the probability $\Pr(y|X)$ of a sentiment y given the features X using the Bayes' Theorem:

$$\Pr(y|X) = \frac{\Pr(X|y)\Pr(y)}{\Pr(X)},$$

We use the naive independence assumption that all features x_i are conditionally independent given y. Hence,

$$\Pr(X|y) = \prod_{i=1}^{n} \Pr(x_i|y).$$

METHODOLOGY

Models: LSTM

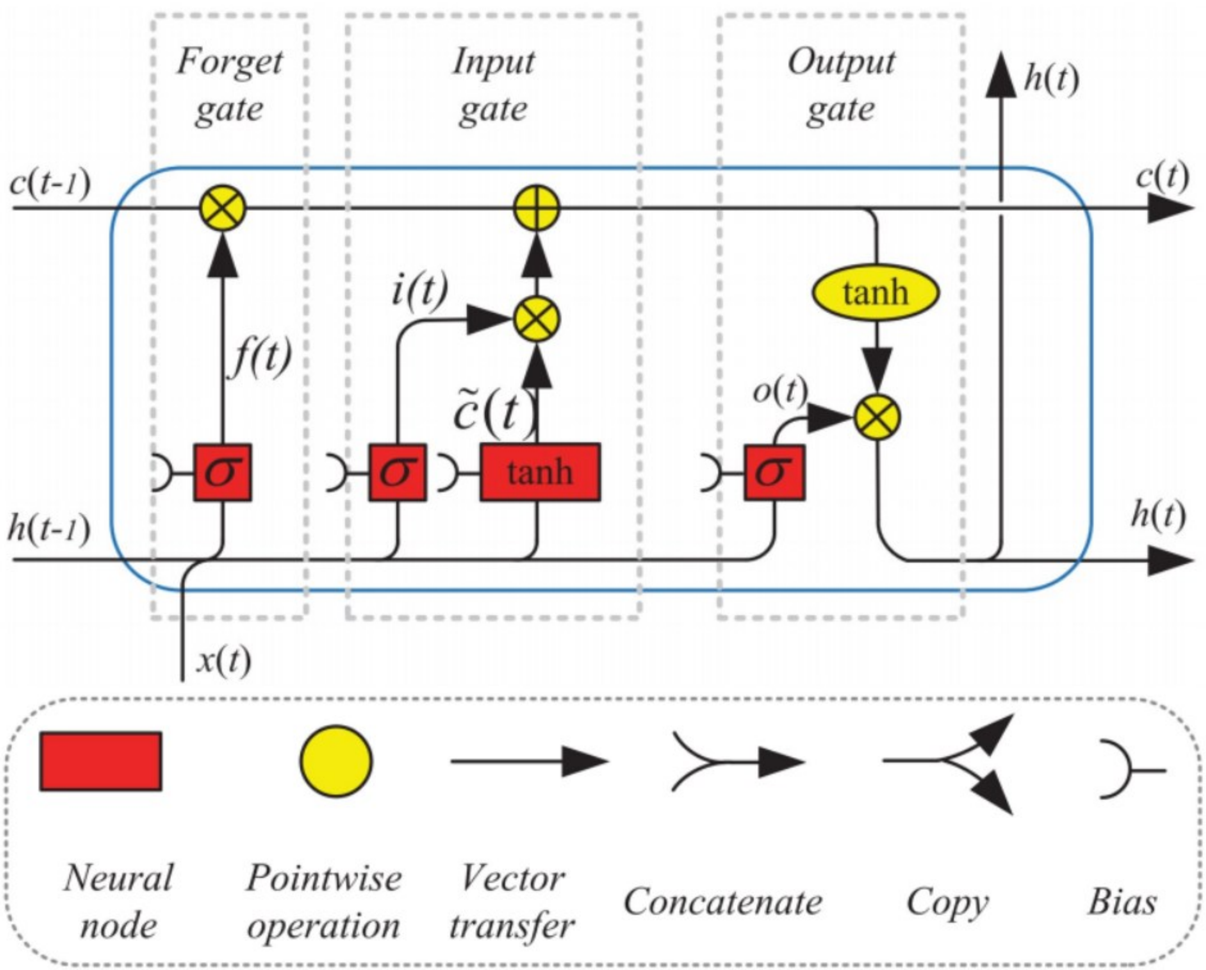


FIGURE 3.6: An LSTM Architecture [63].

- Forget gate.

$$f_t = \sigma(W_f X_t + U_f h_{t-1} + b_f),$$

- Input gate & candidate state.

$$i_t = \sigma(W_i X_t + U_i h_{t-1} + b_i)$$

$$\tilde{C}_t = \tanh(W_c X_t + U_c h_{t-1} + b_c).$$

- Internal cell state

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t,$$

- Output gate & hidden state

$$o_t = \sigma(W_o X_t + U_o h_{t-1} + b_o),$$

$$h_t = o_t \tanh(C_t).$$

S/No	Model	Embeddings	Time to Train[a]	Key Parameters
1	$LSTM_1$	GloVe	1hr 54m	input layer = 100, hidden layer = 128, output = 3
2	LR	DistilBERT	4.63m[b] + 13.4s	l_1 ratio = 0.5
3	NB	Count Vectorizer	2.22s	$\alpha = 0.6$
4	$LSTM_2$	DistilBERT	2.59m[b] + 1hr 20m	input layer = 768, hidden layer = 128, output = 3

[a] This does not account for the time used to train for hyperparameters.
[b] Time to train embeddings on the data using DistilBERT.

TABLE 4.4: Model Training

METHODOLOGY

RabbitMQ

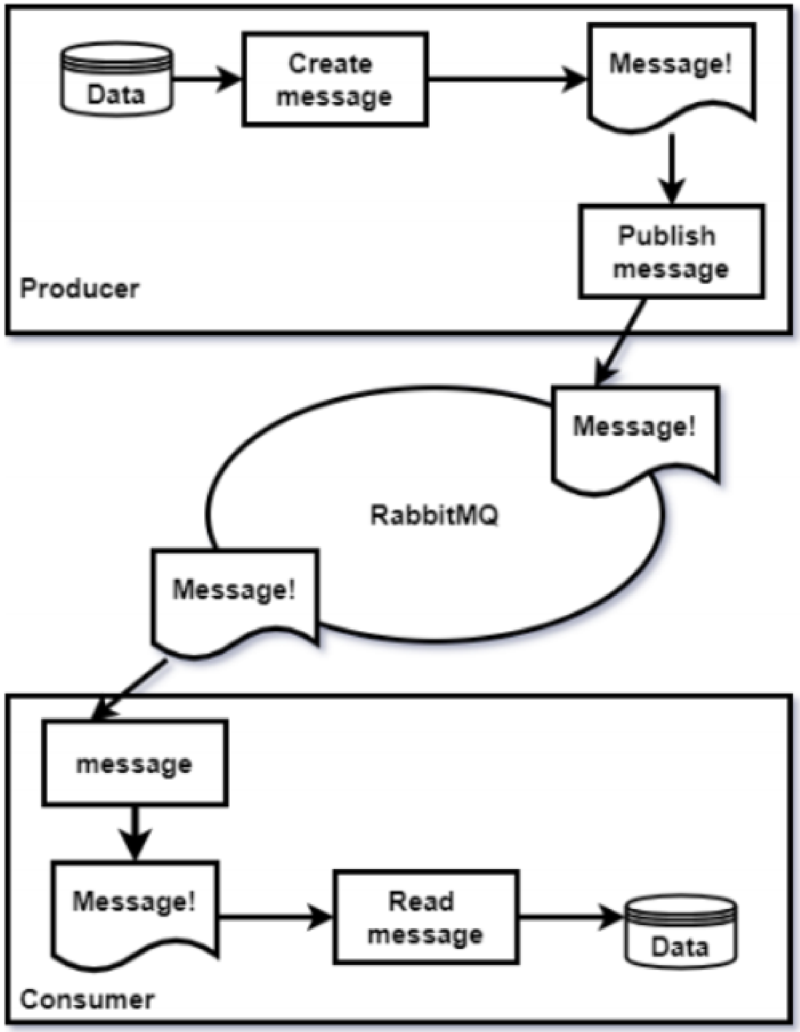
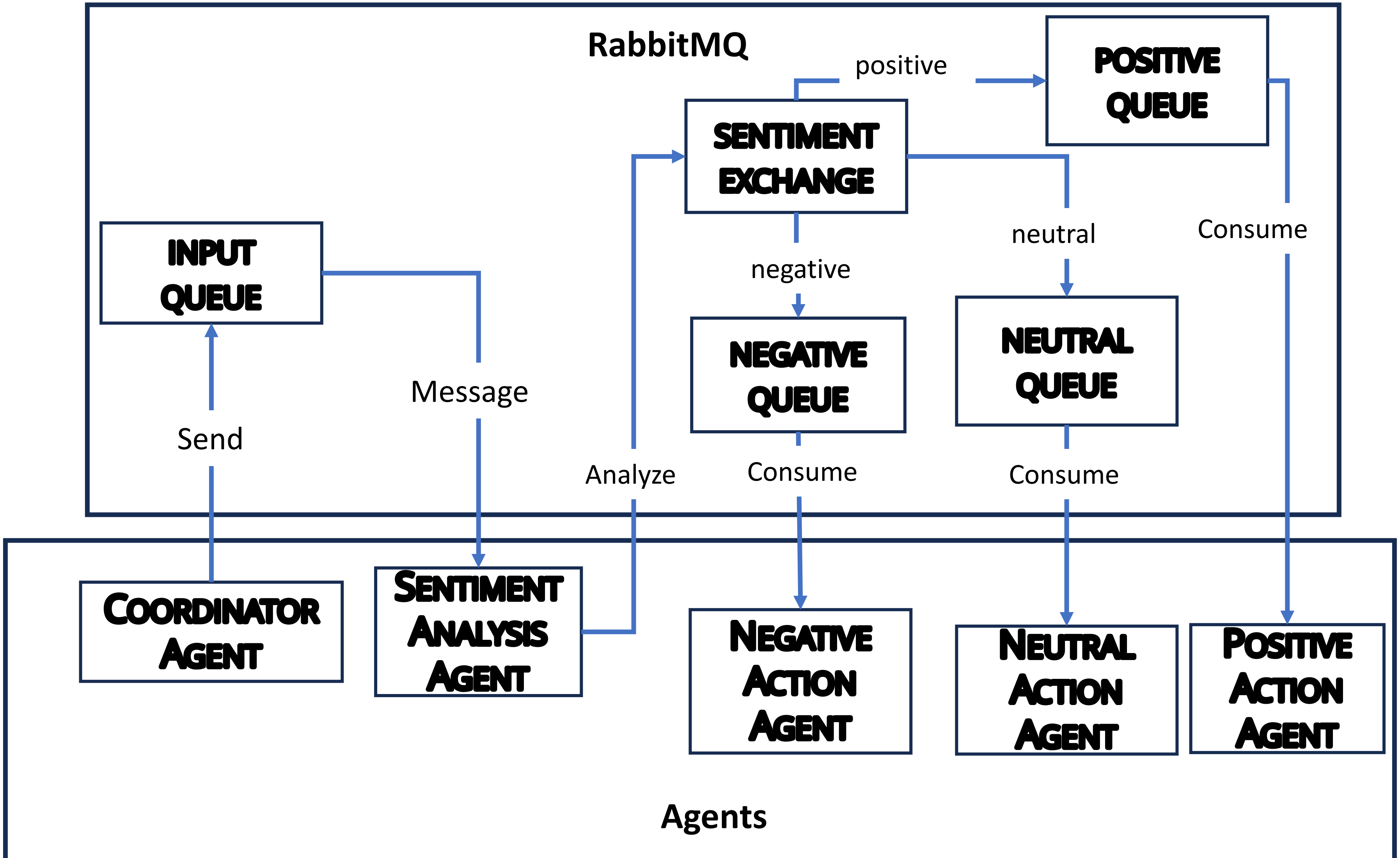


FIGURE 3.2: Message flow from producers to consumers [61].

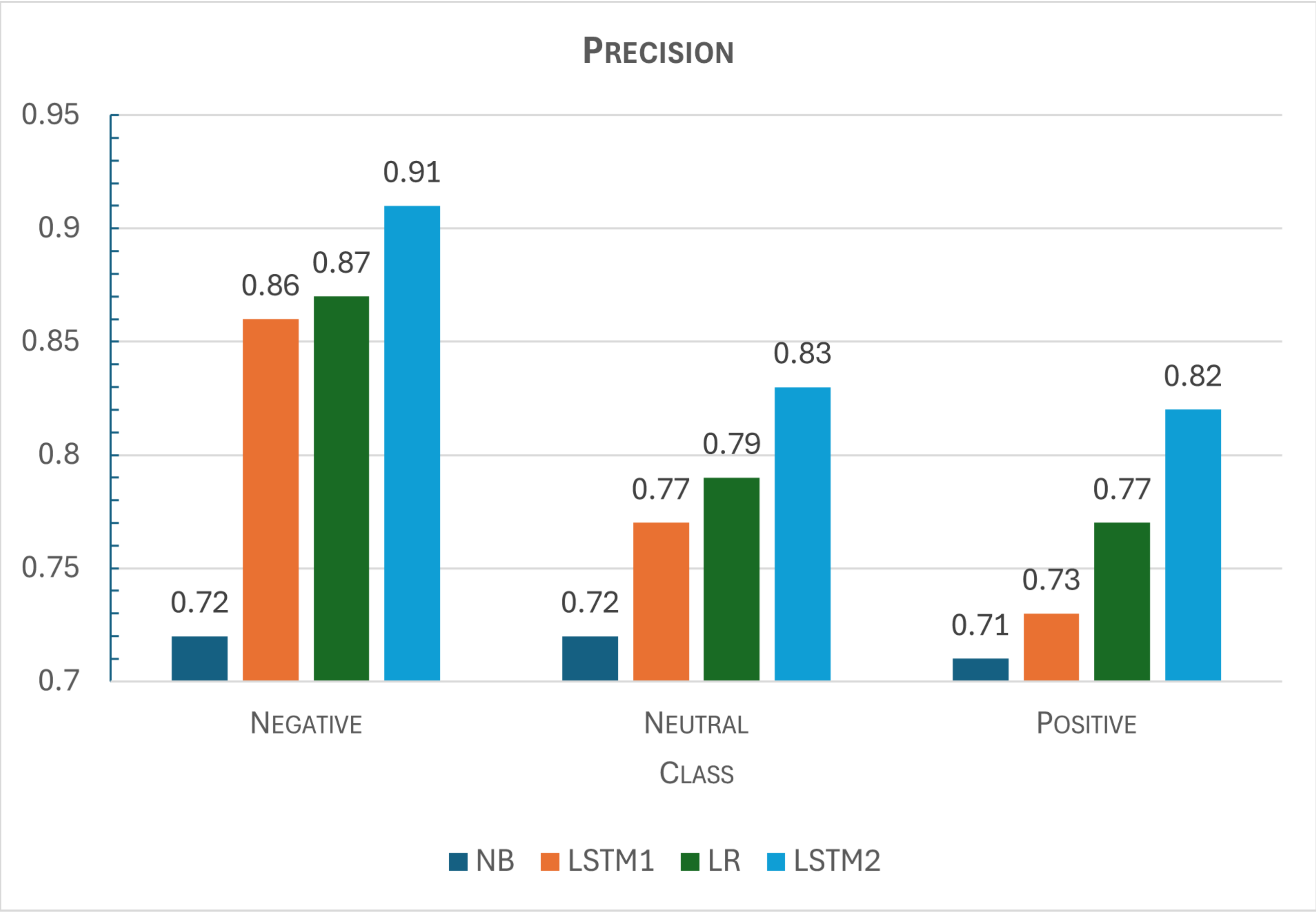
METHODOLOGY

RabbitMQ Messaging Flow



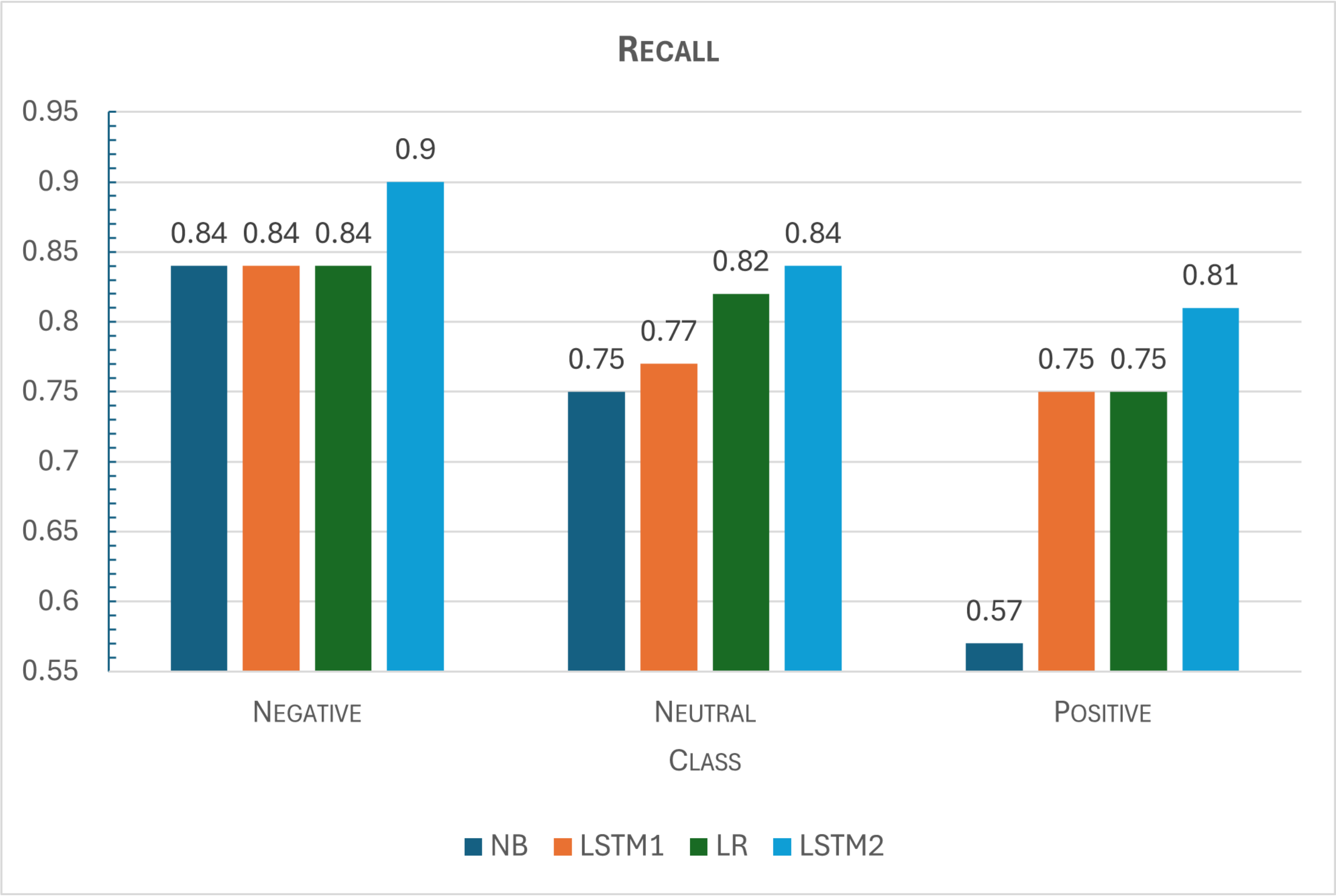
Result

Model Performance

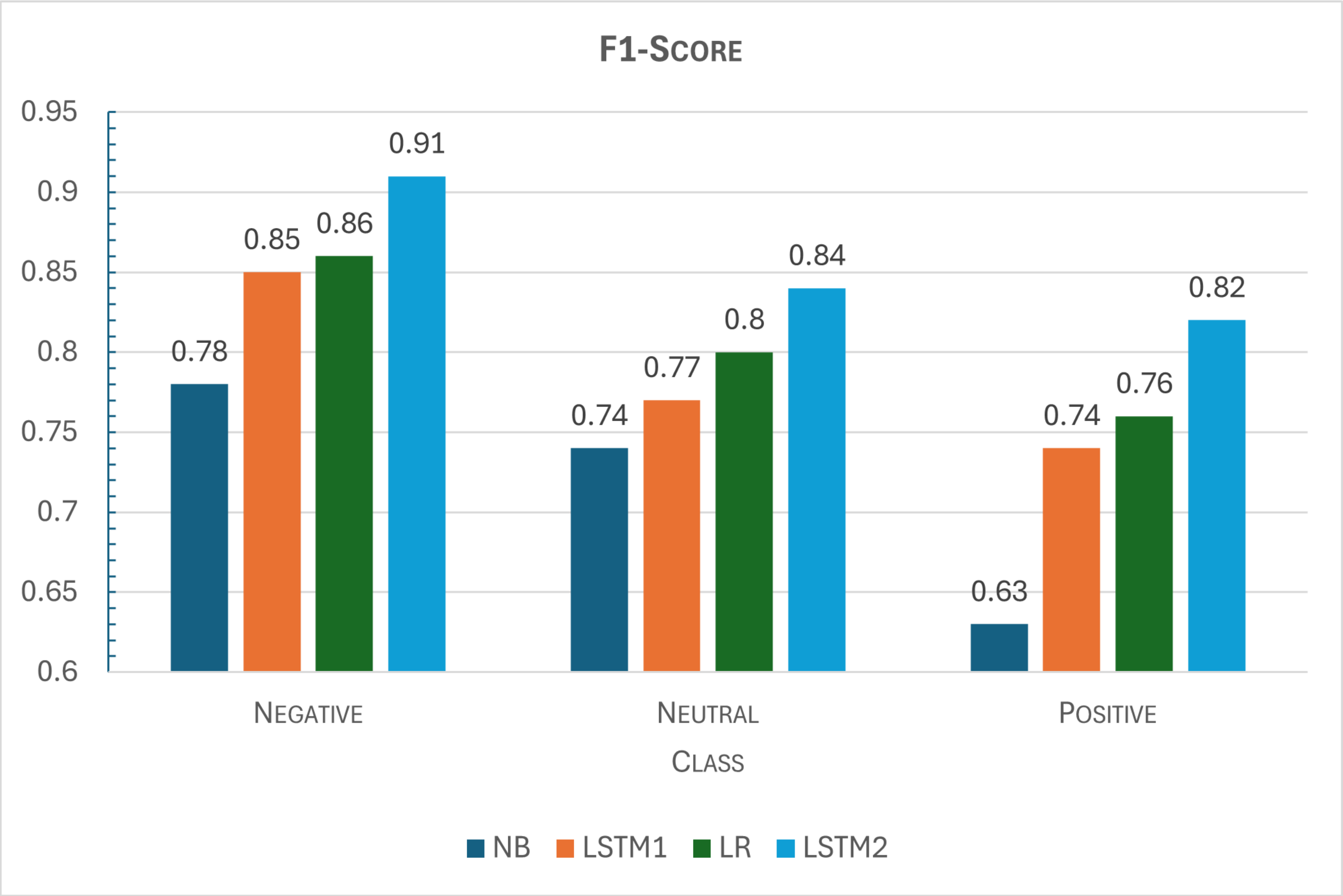


Result

Model Performance



Result
Model Performance

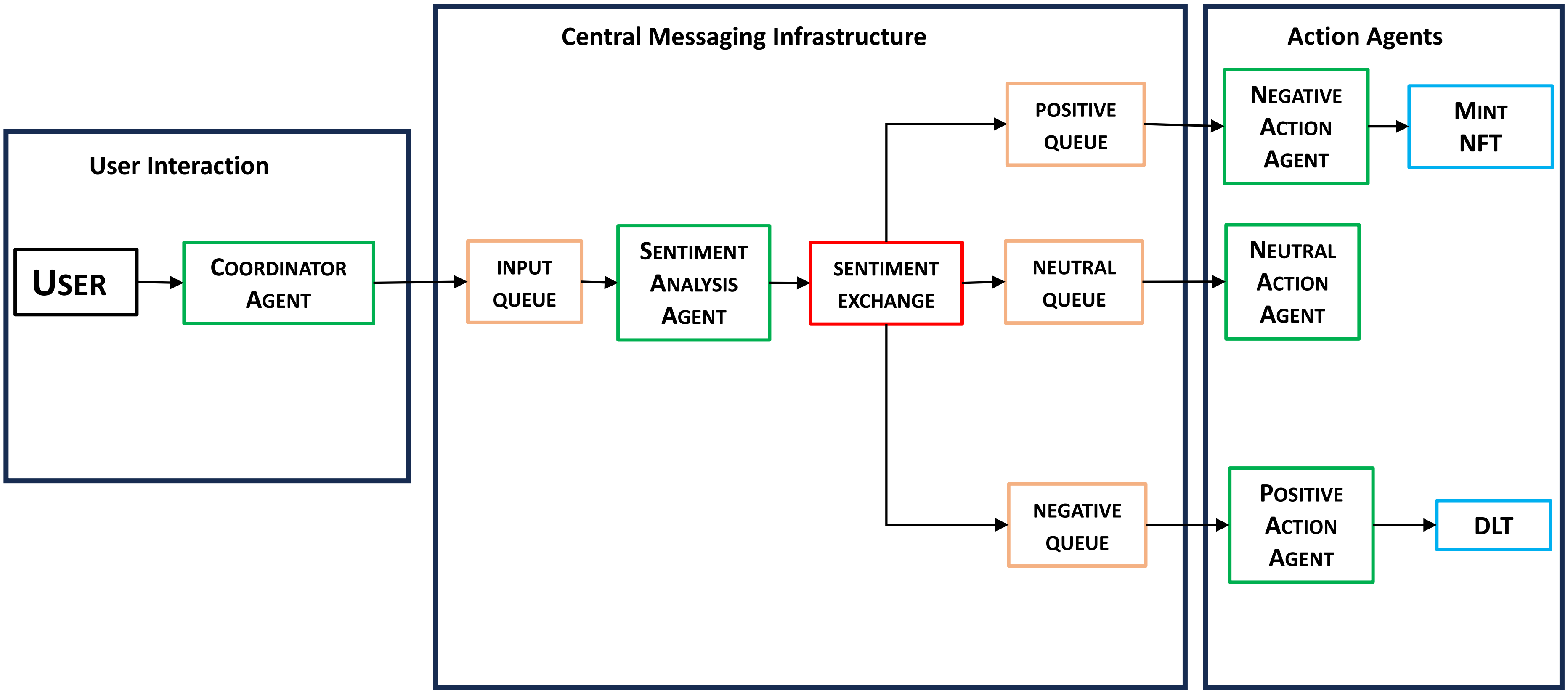


Result

Model Performance

RESULT

MAS Embedded



RESULT

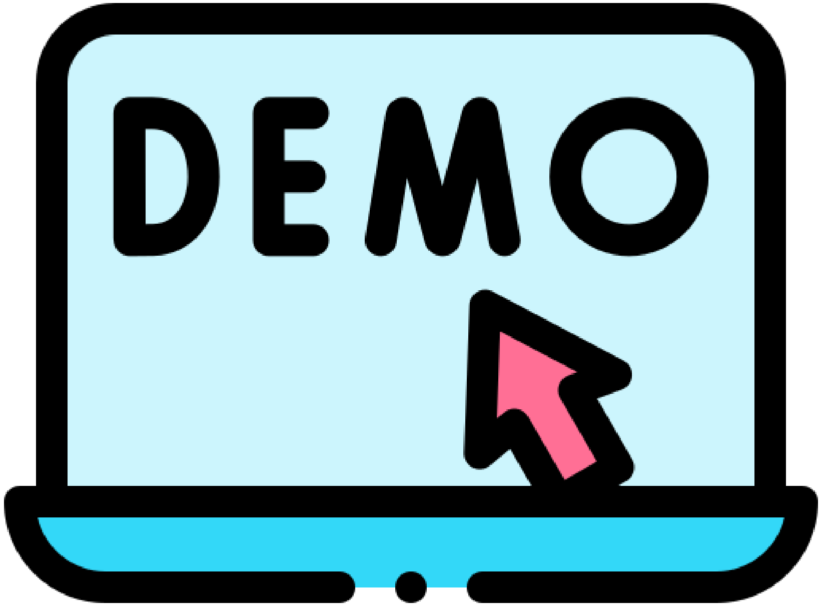
MAS Performance

- Scalable
- Low Communication Latency (fraction of seconds)
- Message Handling

DEMO

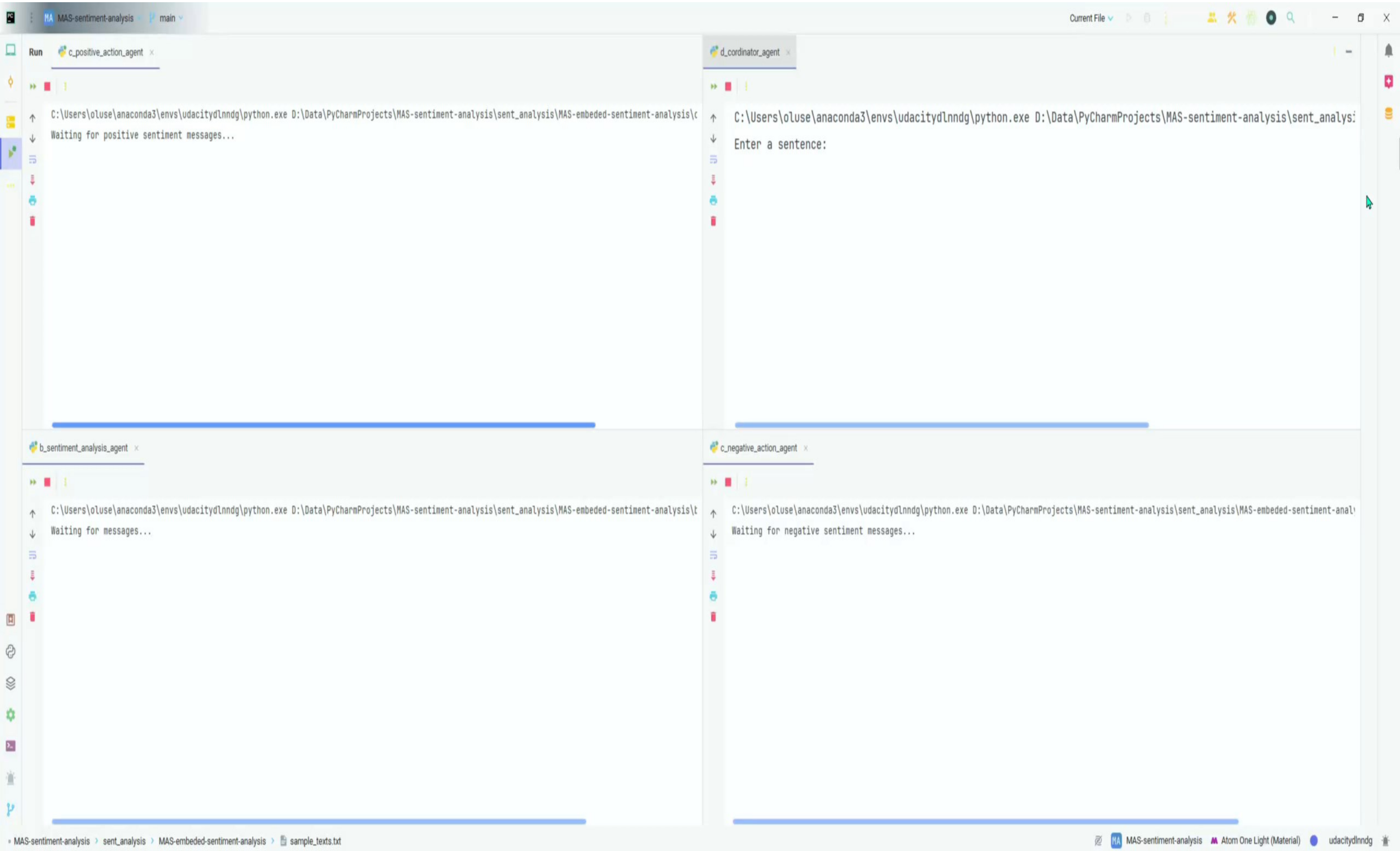
Demonstration of a MAS using RabbitMQ

- The company's stock is making gains in the market despite drastic government policies.
- The stock price tanked after the recent publication of awful earnings for the third quarter.



DEMO

Demonstration of a MAS using RabbitMQ



LIMITATION & FUTURE WORK

- Lack of benchmark financial dataset.
- Absence of GPU for robust training.
- Proprietary financial data.

CONCLUSION

- Design of a MAS
- Models that understand the complexity of financial text
- IOTA ensures transparency, secure and decentralized execution of actions.

Thank you for your attention!