Poem Sentiment Analysis

Bozhi Hao | Xueyi Tan
Quan Yuan | Zhuoyan Guo

30TH/NOV/2023
DSAN 5400

Imagine

Poem

床前明月光，

疑是地上霜。

举头望明月，

低头思故乡。

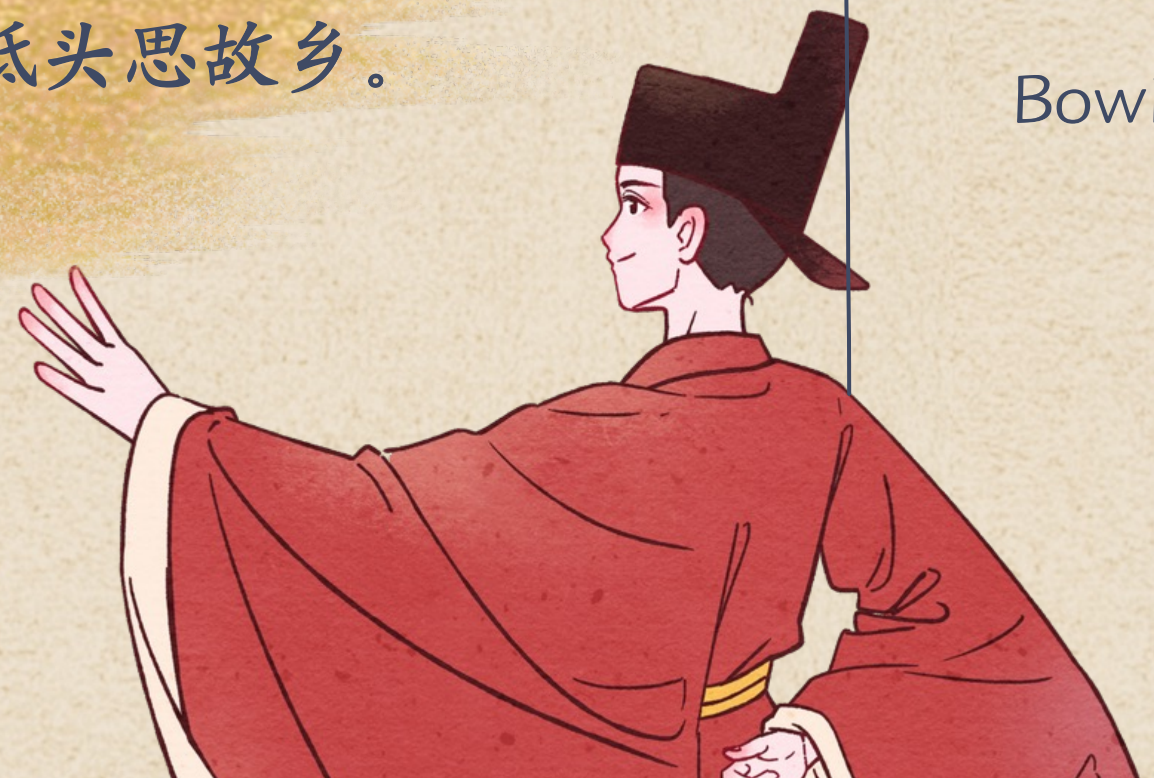
Poem

Afront the bed the Luna beams bright,

Wearing a look of seemingly rime white.

Eyes upcast toward the Luna,

Bowing my head, I think of my hometown



http://127.0.0.1:7860

Imagine

Poem

床前明月光，

疑是地上霜。

举头望明月，

低头思故乡。

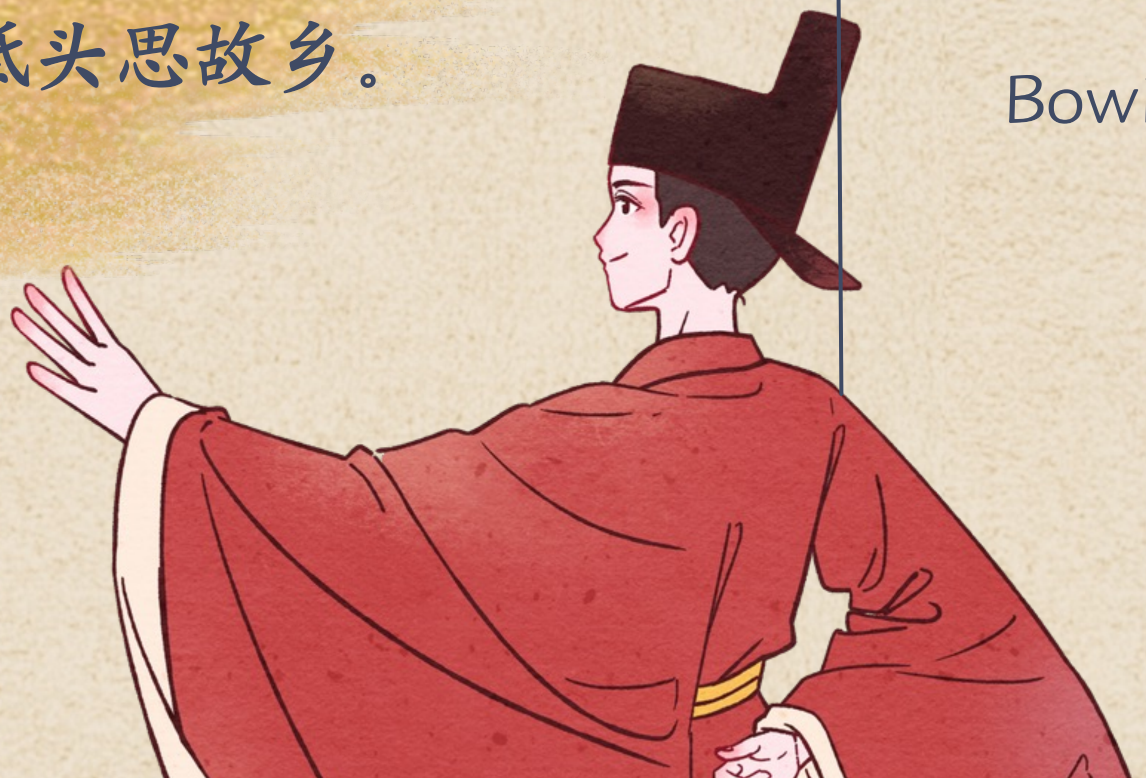
Poem

Afront the bed the Luna beams bright,

Wearing a look of seemingly rime white.

Eyes upcast toward the Luna,

Bowing my head, I think of my hometown



http://127.0.0.1:7860

CONTENTS

Dataset

Sentence	Label
鹅，鹅，鹅，曲项向天歌。白毛浮绿水，红掌拨清波。 Goose, goose, goose. She is singing towards the sky, with her beautiful neck bending in such lovely curve. Her pure white feather coat floats on the Jade green water. er red feet stir crystal wave on the lake surface.	Joy
白日依山尽，黄河入海流。欲穷千里目，更上一层楼。 The white sun behind the mountain falls, The Yellow River into the seas flows.	Worry
天街小雨润如酥，草色遥看近却无。最是一年春好处，绝胜烟柳满皇都。 New grass, visible from afar, vanishing when near. Absolutely the best—before hazy willows becloud the city.	Happiness

Dataset Preprocess

- Split_train_test.py

Emotion	Train data	Test data
Joy	9440	1000
Sadness	13784	1000
Worry	3977	300
Thoughtfulness	10550	1000
Happiness	6578	500
Anger	2158	200
Fear	493	100

Dataset Preprocess

Convert the dataset into vectorized data

Vocab class：

1. fit: iterate over sentences to count word occurrences
2. build_vocab: applies our frequency thresholds to finalize the vocabulary
3. transform: convert sentences to their numerical representations
4. inverse_transform: To interpret model outputs or for verification, we can reverse the numerical sequences back to words using the inverse dictionary we created during vocabulary building

```
低头思故乡
['低', '头', '思', '故', '乡']
[1039, 409, 152, 1027, 938, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
sad
```

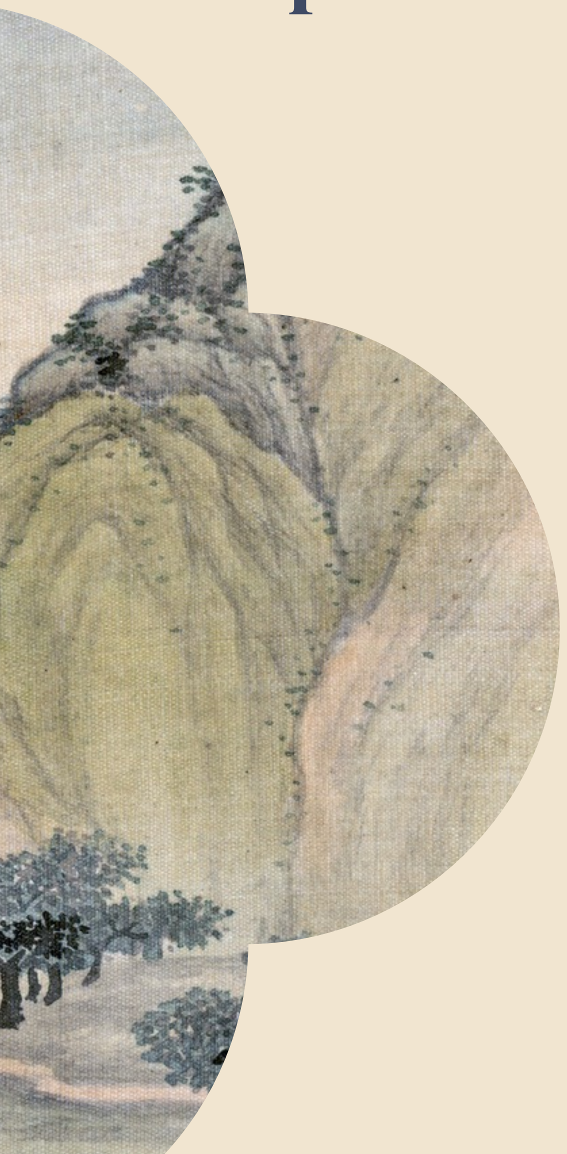
Attention Mechanism

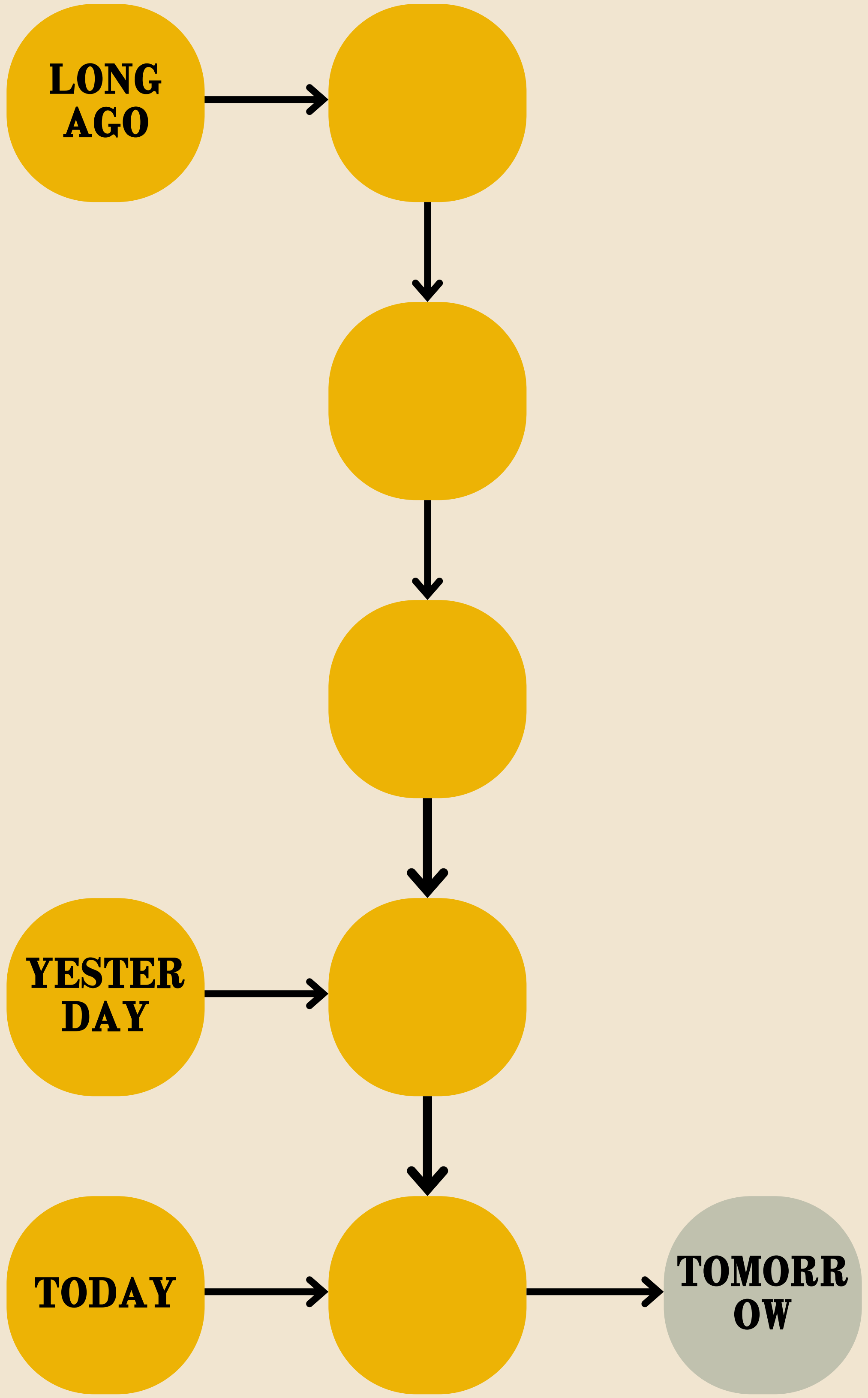
Attention Mechanism

- A mechanism mimicking cognitive attention
- Pays greater attention to certain factors when processing the data
- In our project, it helps the system more accurately capture the key words or phrases that express the sentiment in the poems

Introducing awesome Attention Mechanism

- Recurrent Neural Networks had problems with long-term memories because they ran both the long and short-term memories through a single path...

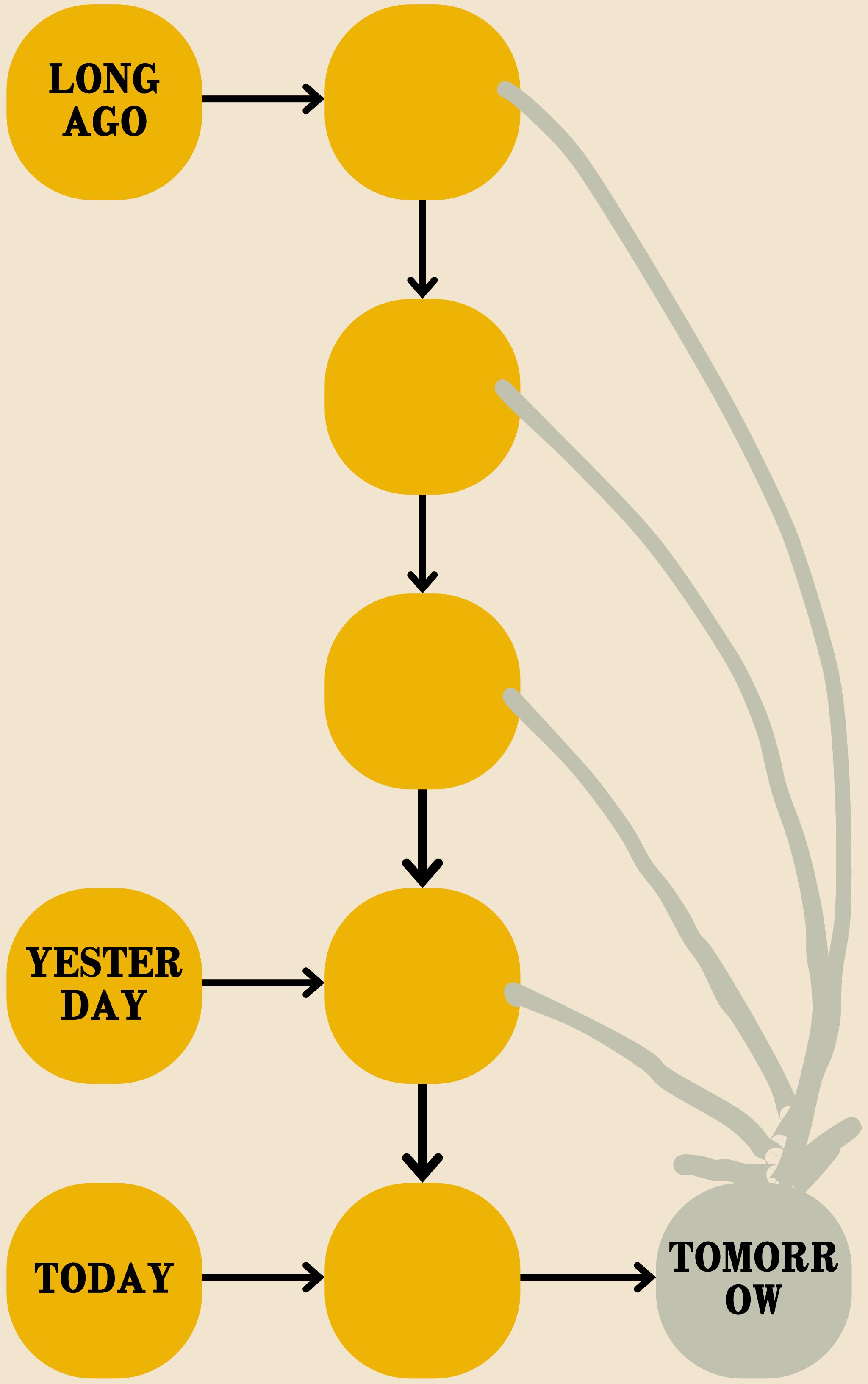




Introducing awesome Attention Mechanism

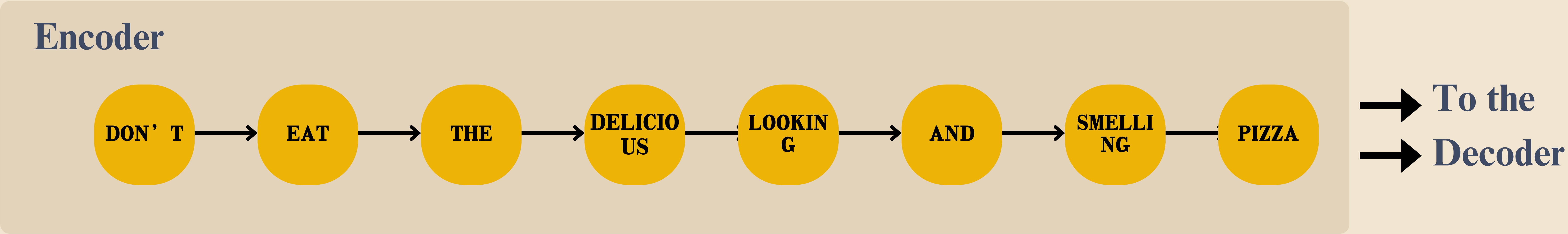
- Recurrent Neural Networks had problems with long-term memories because they ran both the long and short-term memories through a single path...

- LSTM solve this problem by providing separate paths for long & short term memories



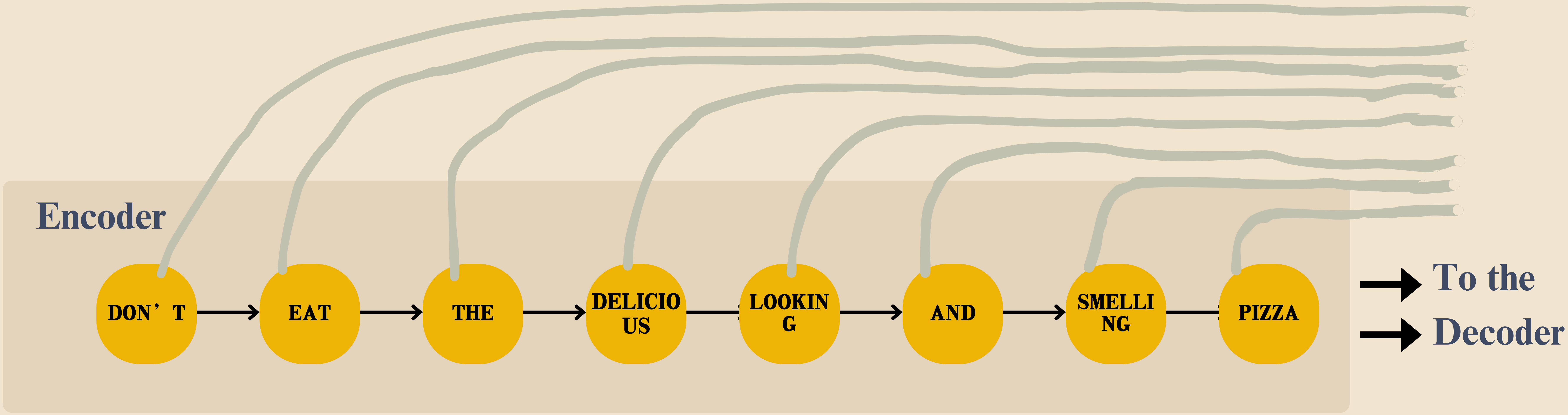
Introducing awesome Attention Mechanism

- Even with LSTMs words, with separate paths, if we have a lot of data, both paths have to carry a lot of information, words that are input early on can be forgotten
- And the meaning will be reversed!



The main idea of Attention

- So the Main Idea of Attention is to add a bunch of new paths from the Encoder to the Decoder, one per input value, so that each step of the Decoder can directly access input values

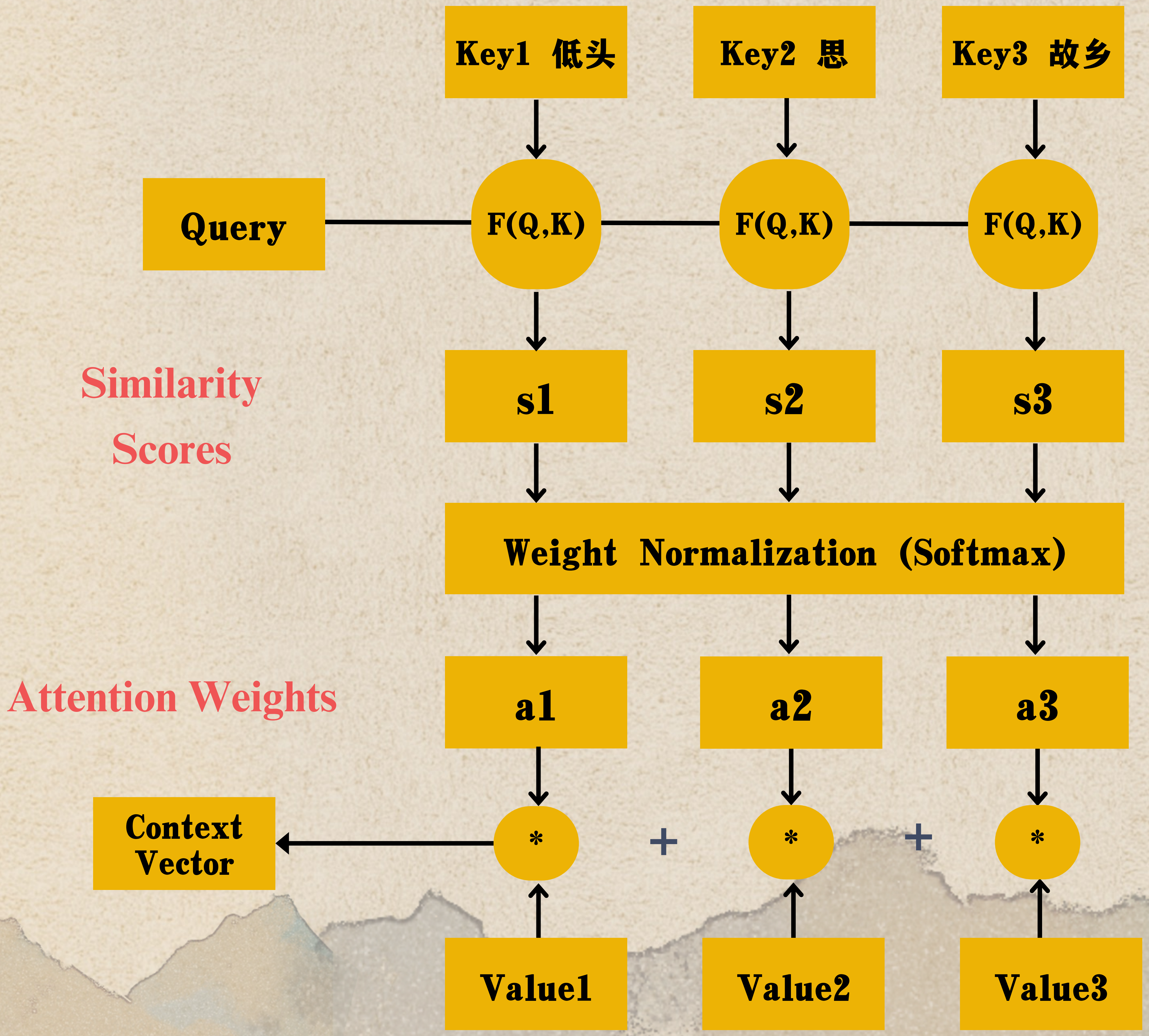


The main idea of Attention

Poem

低头思故乡

Bowing my head, I think of my hometown

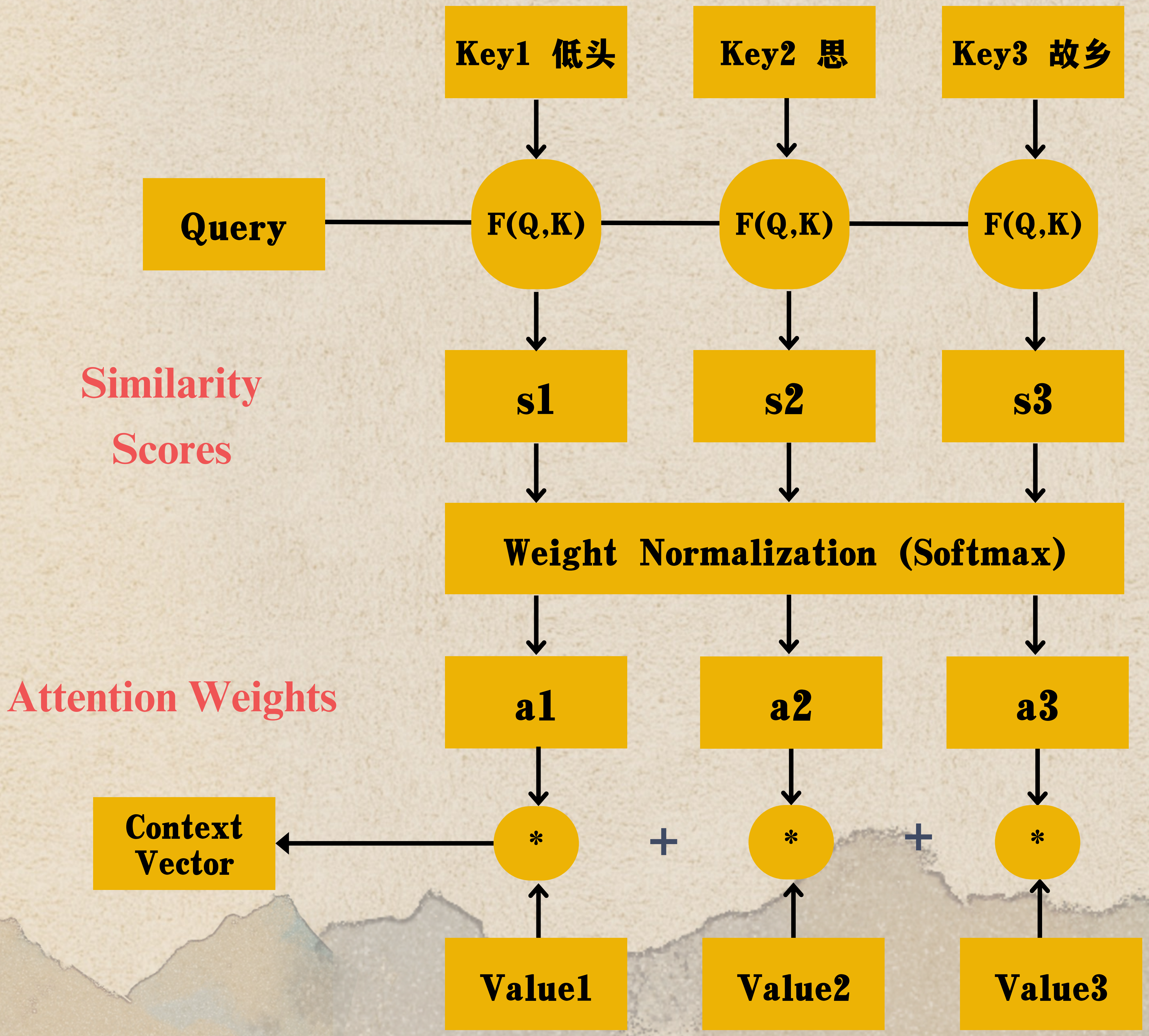


The main idea of Attention

低头思故乡

Bowing my head, I think of my hometown

- Input Data and Encoder Output (X=[x1, x2, ..., xn])
- Compute similarity scores between each key and the Query Vector
- Normalize scores with softmax: adding up to 1
- Calculate the context vector with value and attention weights



BiLSTM

Why BiLSTM

- Contextual Information: BiLSTM considers both preceding and following text, aiding in understanding and capturing emotional nuances in context.
- Handling Long-Distance Dependencies: capture long-term dependencies is valuable for accurately analyzing emotions.
- Polysemy Resolution: BiLSTM's ability to capture multiple contextual nuances from different directions aids in resolving polysemy issues.

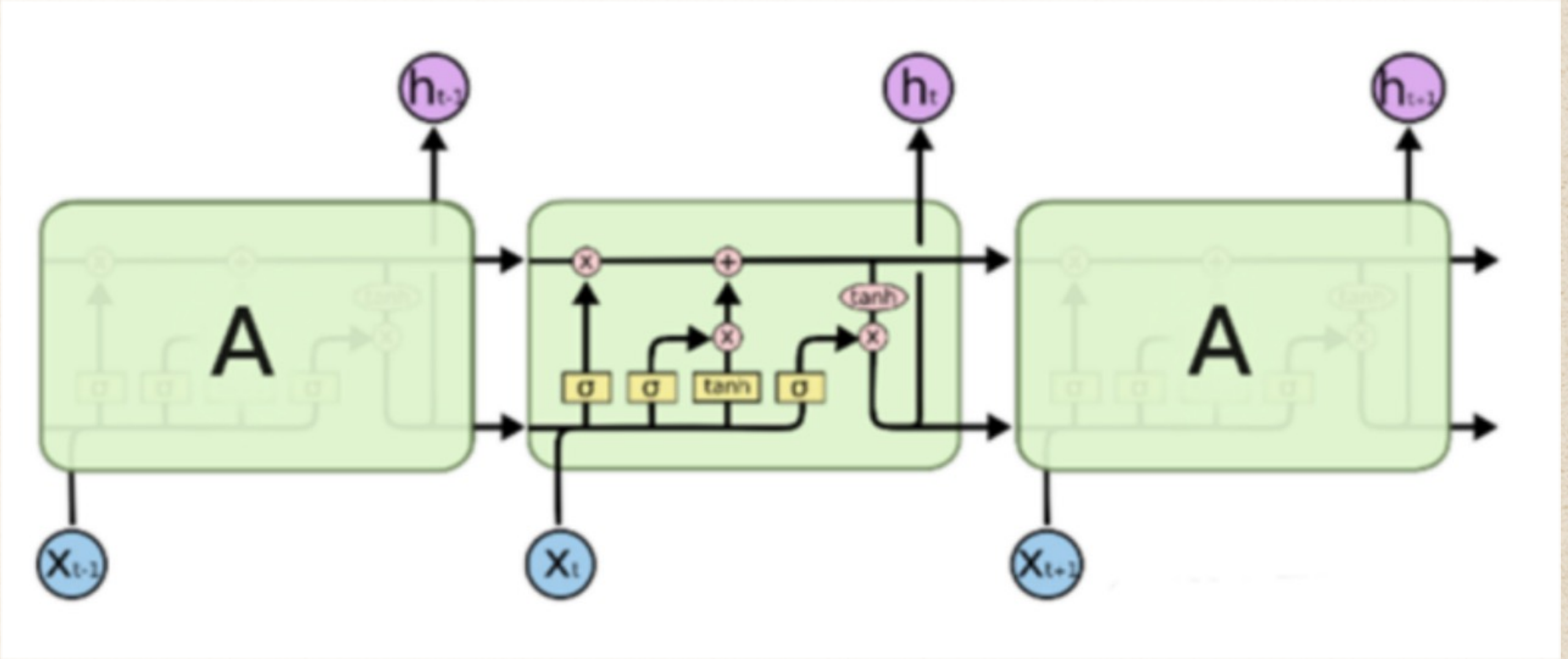
The main process of BiLSTM

BiLSTM is contructed by three gates

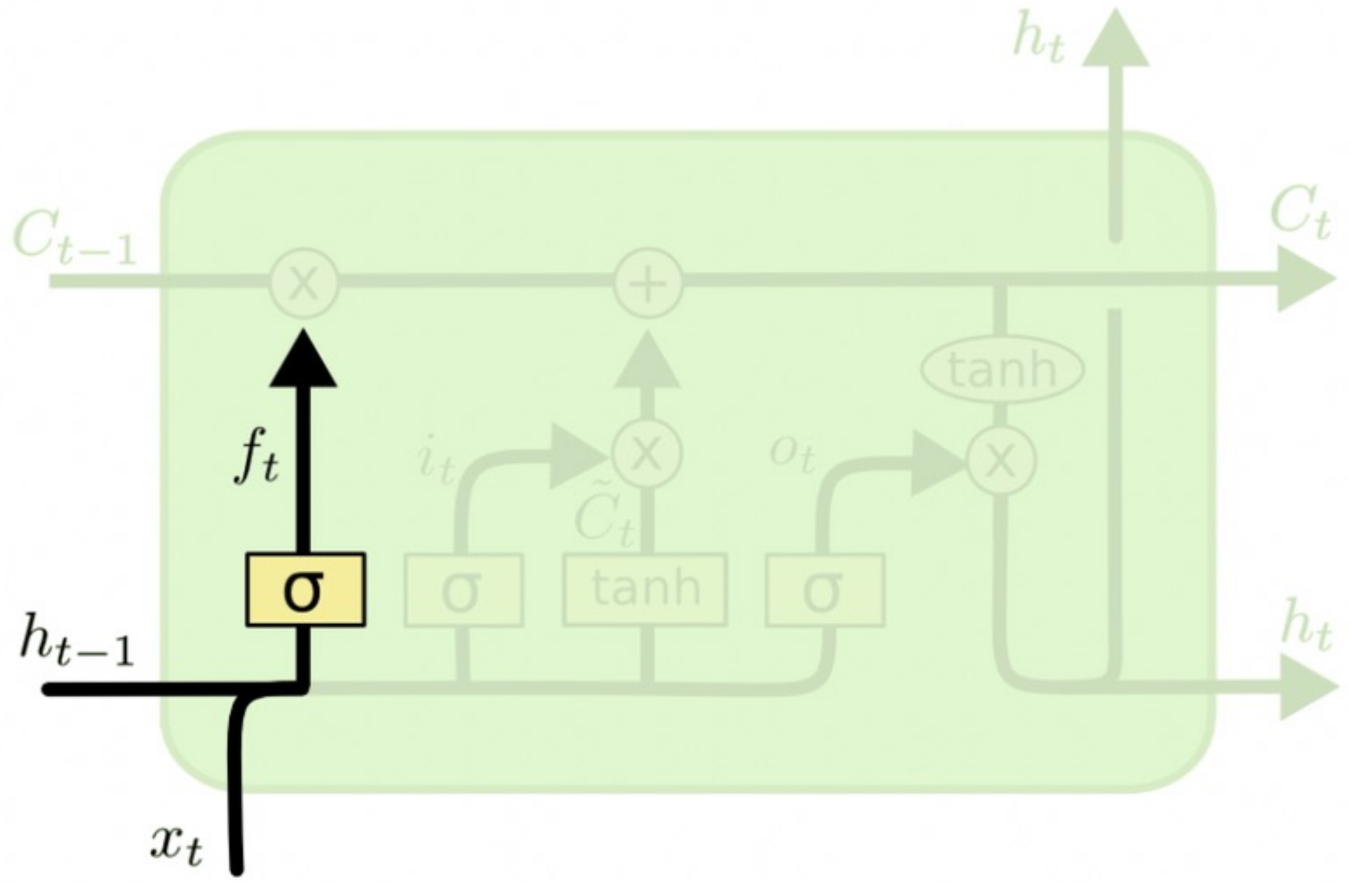
1.Forget gate

2.Input gate

3.Output gate

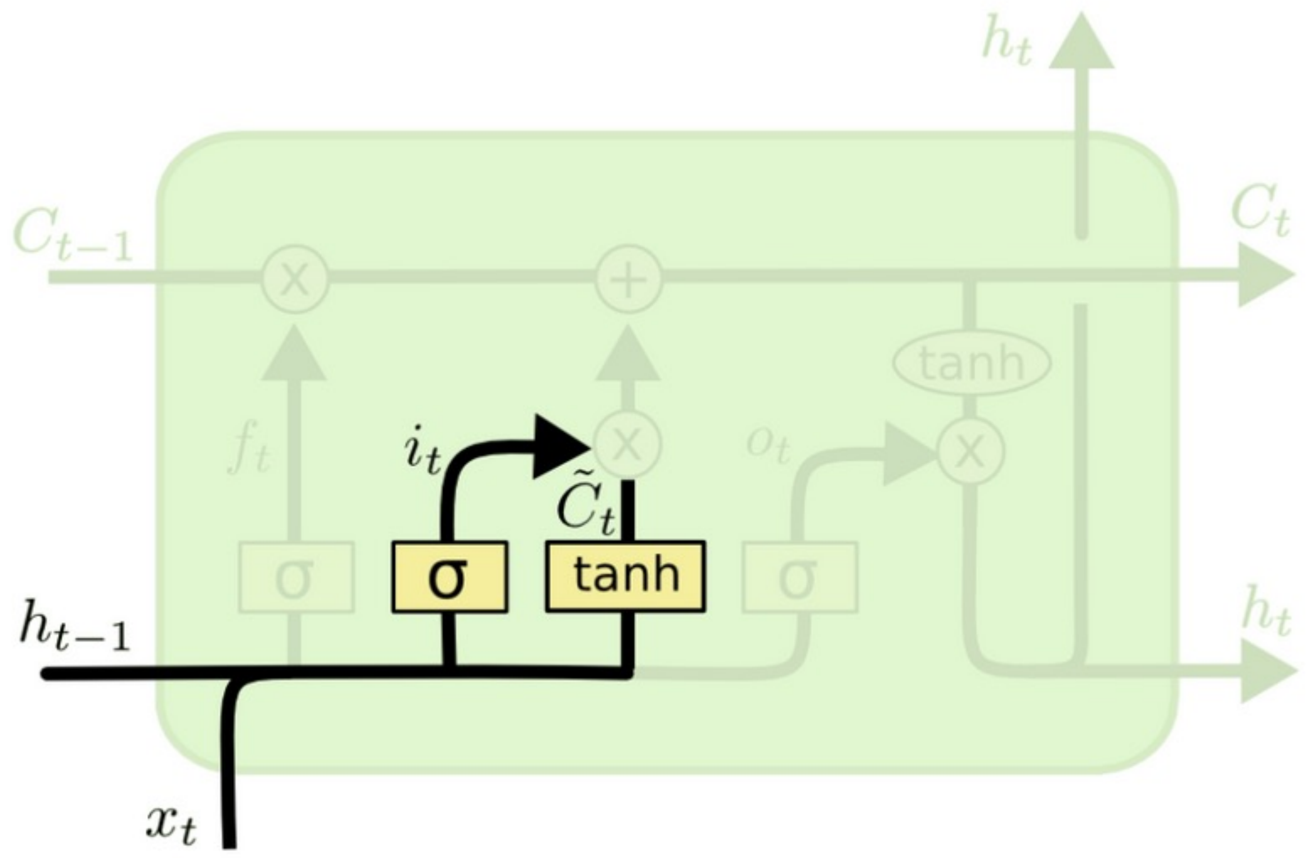


Forget Gate



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

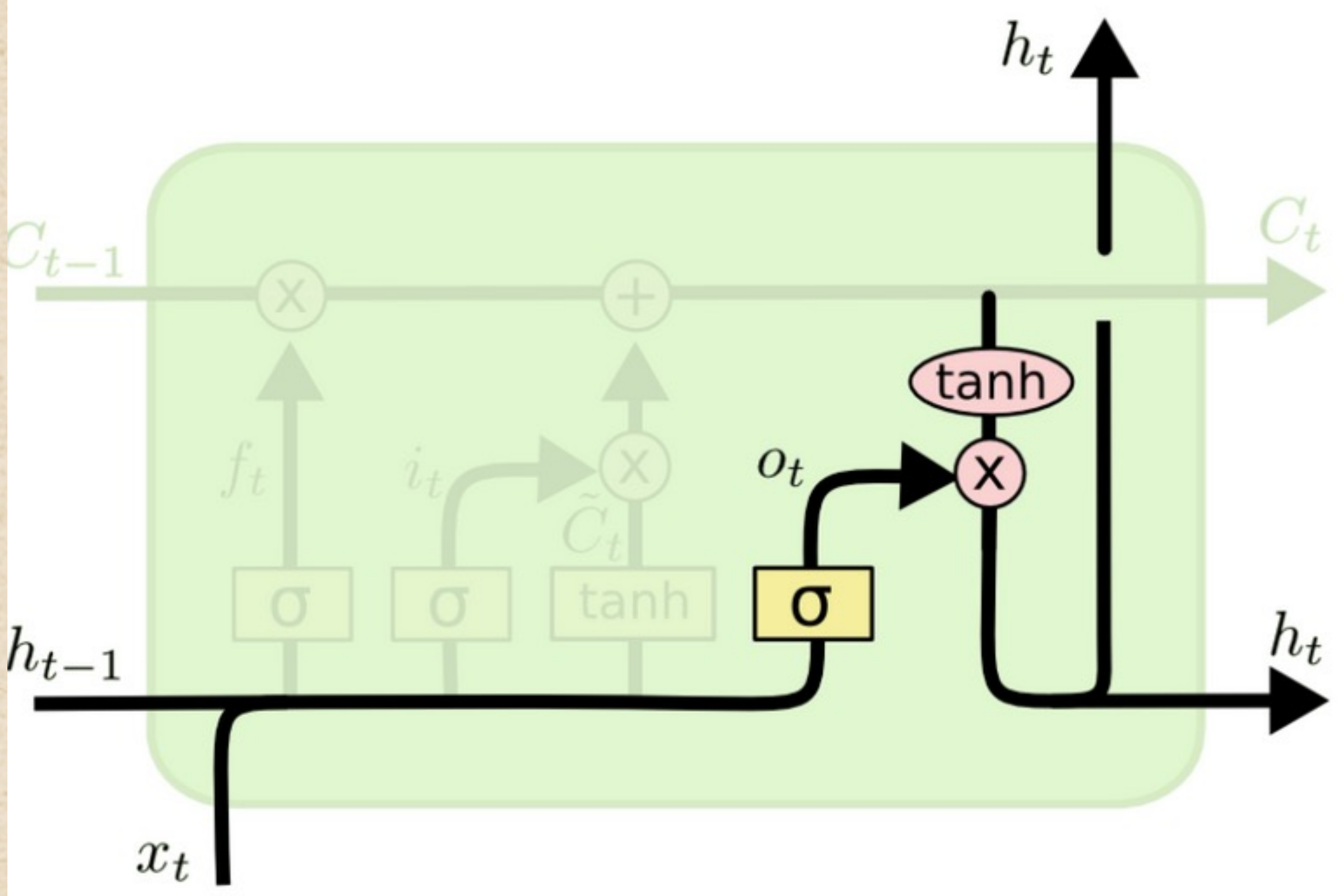
Input Gate



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

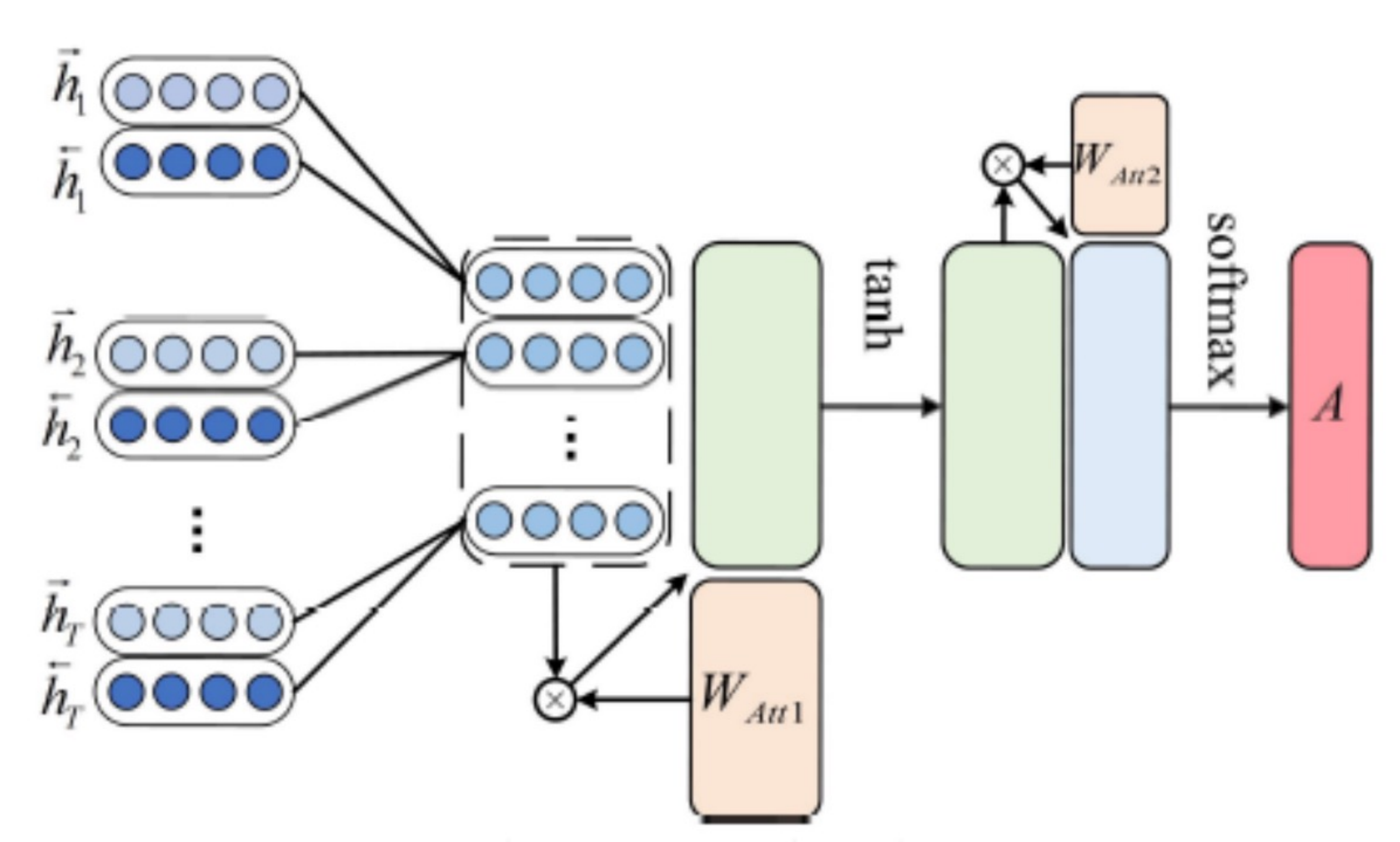
Output Gate



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Merge Layer


$\vec{h}_1$
$\overleftarrow{h}_1$
$\vec{h}_2$
$\overleftarrow{h}_2$
$\vec{h}_T$
$\overleftarrow{h}_T$
W_{Att1}
tanh
W_{Att2}
softmax
A

Output Layer

The model extracts and fuses information from the sentence, resulting in high-dimensional matrices with emotional factors. However, these matrices don't yield sentiment classifications directly. To achieve this, a fully connected layer is added to map them into sentiment classifications for poetic lines which is designed as class ImdbModel in our project.

Conclusion

第四

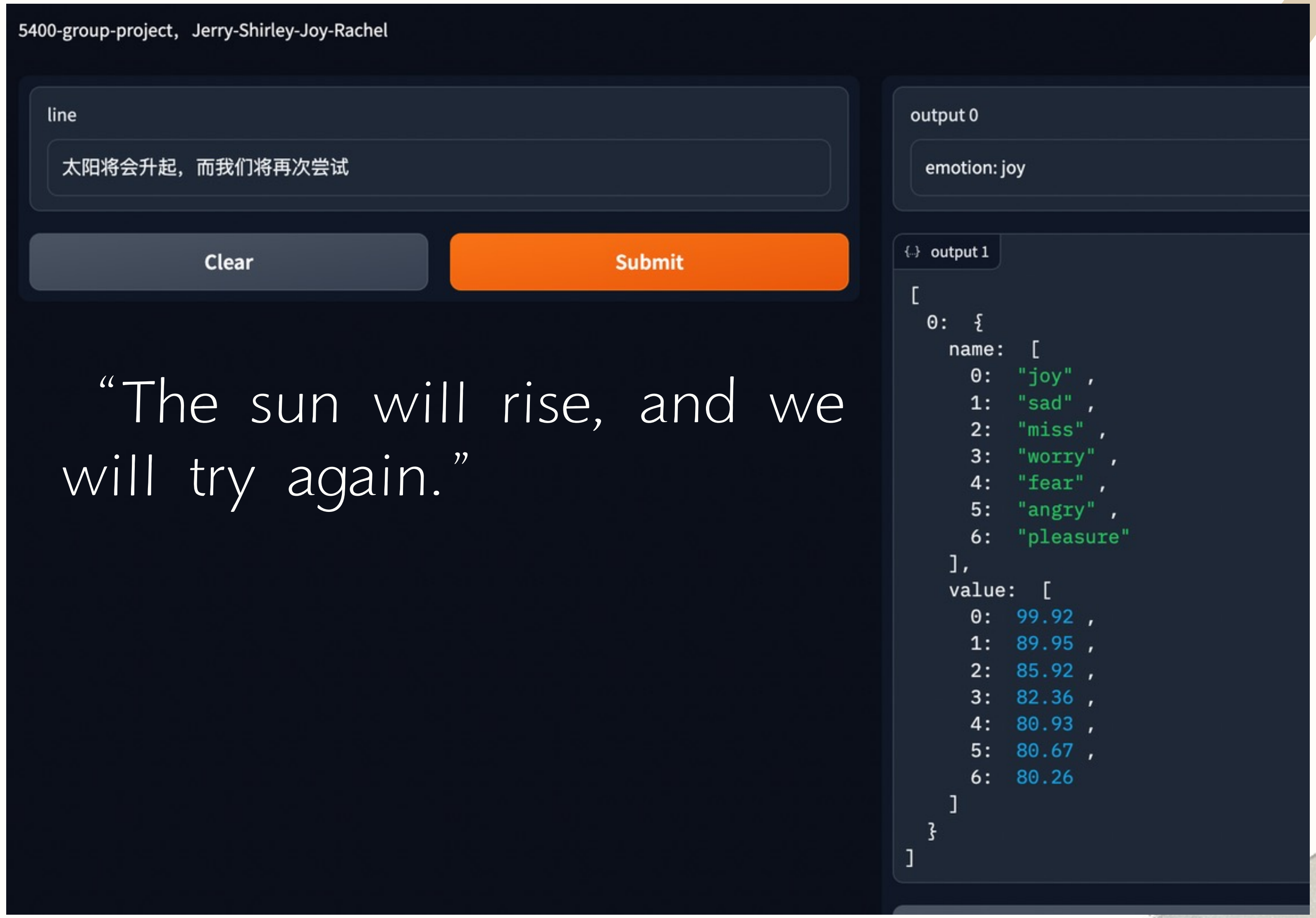
Designing the User Interface

Model Integration

Incorporated the trained emotion classification model (lstm_model.pkl) into the interface.

GUI

Utilized common Graphical User Interface tools for a user-friendly experience



Data Preprocessing

Same preprocessing steps as the training data, to ensure consistency

Classification Prediction

The poetry text is classified into emotional categories based on the model’s output

Performance Metrics

Accuracy	0.7558	the ratio of the number of samples correctly predicted by the model to the total number of samples.
Precision	0.7184	the proportion of samples that the model predicts to belong to a certain emotion category that actually do belong to that category.
Recall	0.6619	the ratio of the number of samples that the model correctly predicts to belong to a certain emotion category to the number of samples of that category in the test set
F1 Score	0.6757	integrates both the accuracy and the recognition capability of the model and is an important metric for evaluating model performance

第四章

Limitations and Improvements

- Data Dependence
  - Sadness：13,784
  - Fear：493
  - More Varied & Extensive Dataset
- Improving Interpretability
  - Attention Mechanism Limitations
  - Visualization Tools; Rule-Based Explanation Techniques
- Word-Based Tokenization
  - Phrase-Based Tokenization

Reference

Sreeja, P. S., & Mahalakshmi, G. S. (2016). Comparison of probabilistic corpus based method and vector space model for emotion recognition from poems. Asian Journal of Information Technology, 15(5), 908-915.

Hou, Y., & Frank, A. (2015, July). Analyzing sentiment in classical Chinese poetry. In Proceedings of the 9th SIGHUM workshop on language Technology for Cultural Heritage, social sciences, and humanities (LaTeCH) (pp. 15-24).

Zehe, A., Becker, M., Jannidis, F., & Hotho, A. (2017). Towards sentiment analysis on German literature. In KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI, Dortmund, Germany, September 25 – 29, 2017, Proceedings 40 (pp. 387-394). Springer International Publishing.

Kaur, J., & Saini, J. R. (2017, February). Punjabi poetry classification: the test of 10 machine learning algorithms. In Proceedings of the 9th international conference on machine learning and computing (pp. 1-5).

Jareanpon, C., Kiatjindarat, W., Polhome, T., & Khongkraphan, K. (2018, January). Automatic lyrics classification system using text mining technique. In 2018 international workshop on advanced image technology (IWAIT) (pp. 1-4). IEEE.

THANKS