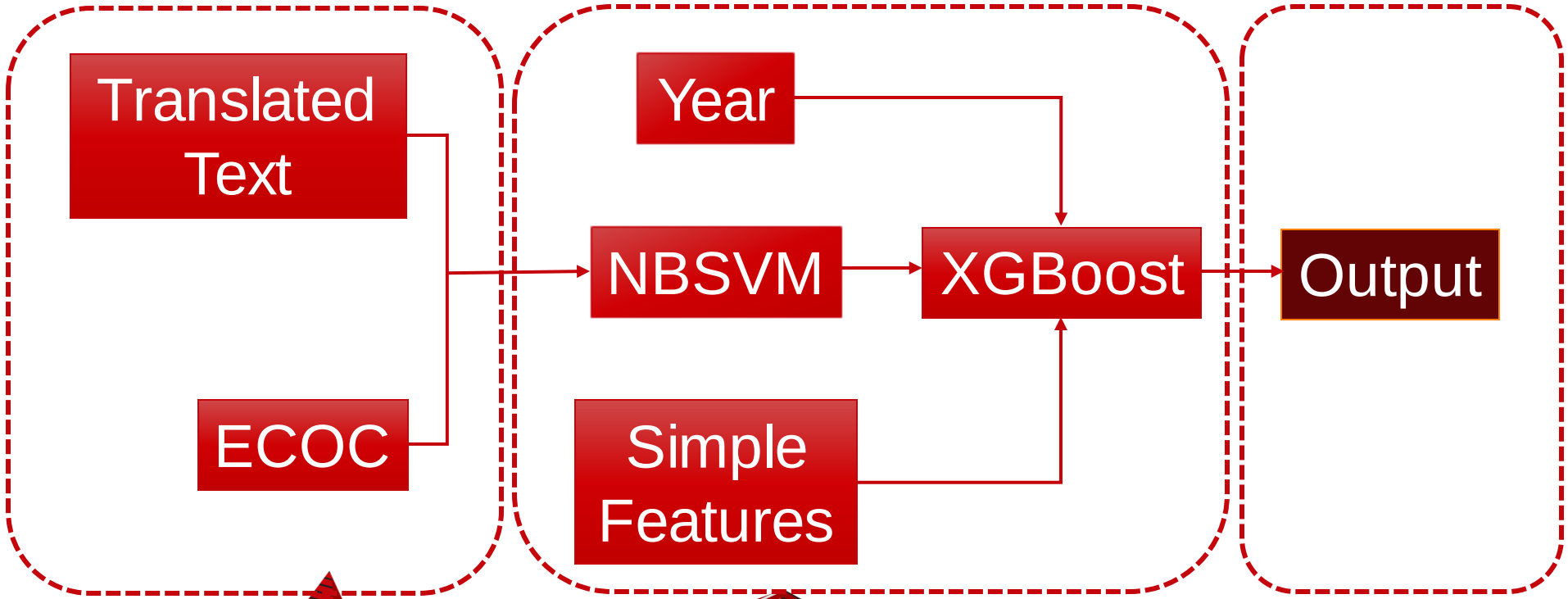
Yelp Ratings Prediction

Presentation 2

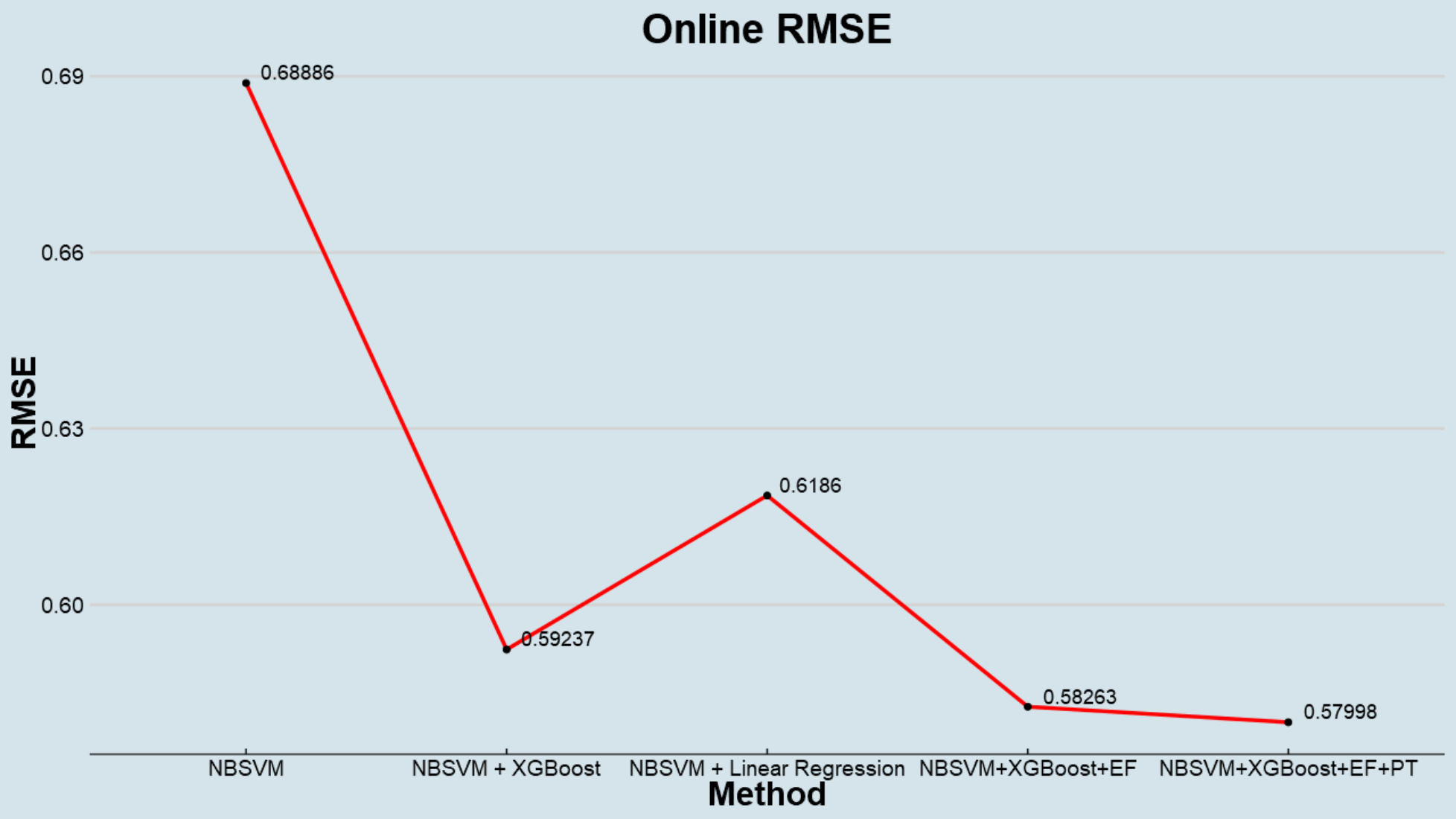
Group 6: Qizheng Ren, Hanmo Li, Lixia Yi, Jiacheng Xu

Introduction

- Public Leaderboard Score: 0.57998
- Model Process:



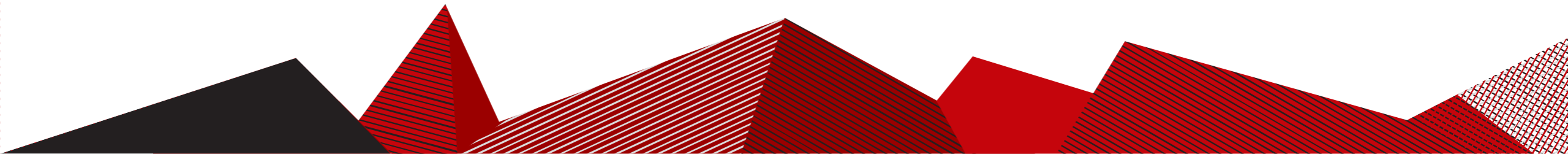
Timeline


Online RMSE
RMSE
0.69
0.66
0.63
0.60
0.68886
0.59237
0.6186
0.58263
0.57998
NBSVM
NBSVM + XGBoost
NBSVM + Linear Regression
NBSVM+XGBoost+EF
NBSVM+XGBoost+EF+PT
Method

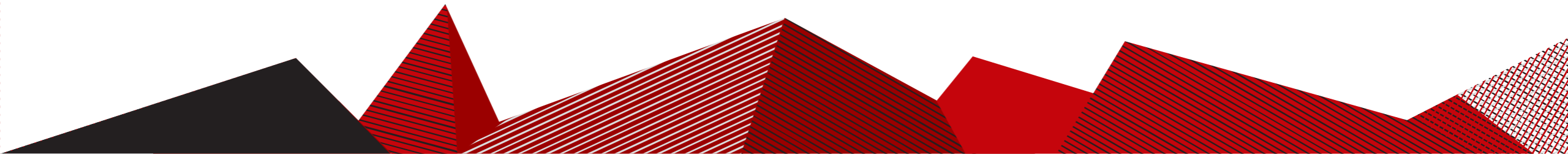
Data Cleaning: Translation

- More than 99% of the reviews were written in English
- Other languages include French, Japanese, German, Chinese
- Most of them are well written but not necessarily translated well

3 - The interior decoration is very beautiful, things out too slow, the food is very beautiful, but things like the general! Ten dollars on the price a bit too expensive. Will I be too picky!

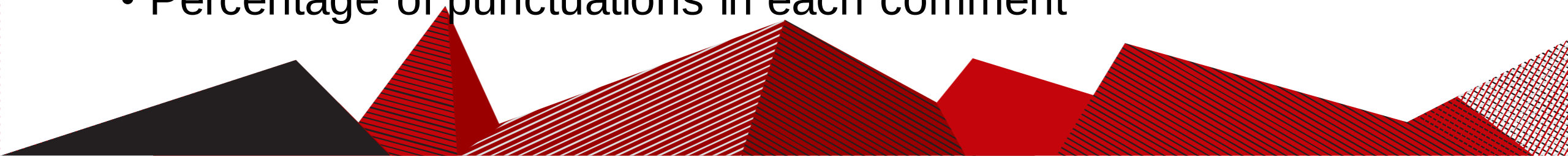


And that's all for cleaning...

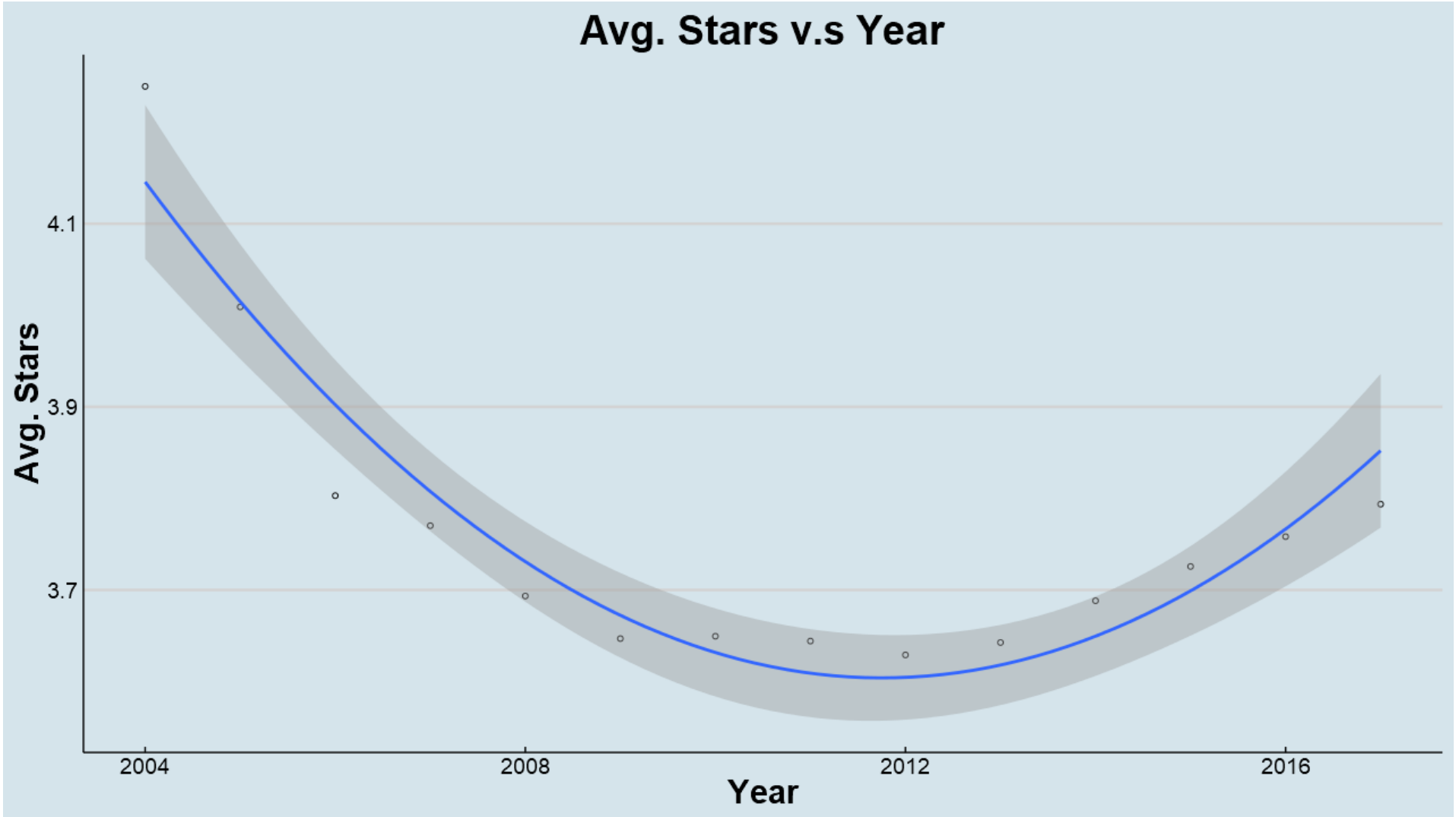


Extra features (EF)

- The number of sentences in each comment
- Word count in each comment
- Unique word count
- Letter count
- Punctuation count
- Upper case words count
- Title case words count
- Number of stopwords
- Average length of the words
- Percentage of unique words in each comment
- Percentage of punctuations in each comment



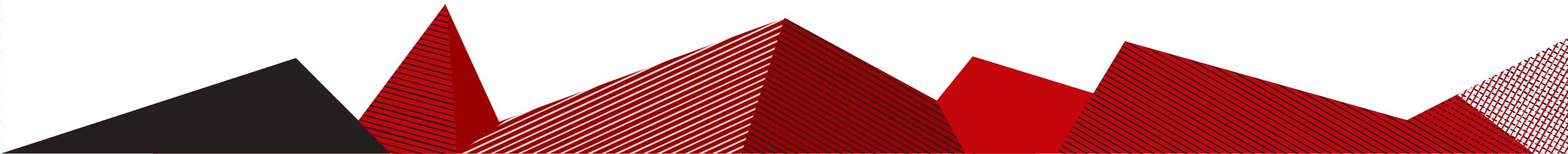
Years



Error-Correcting Output Coding (ECOC)

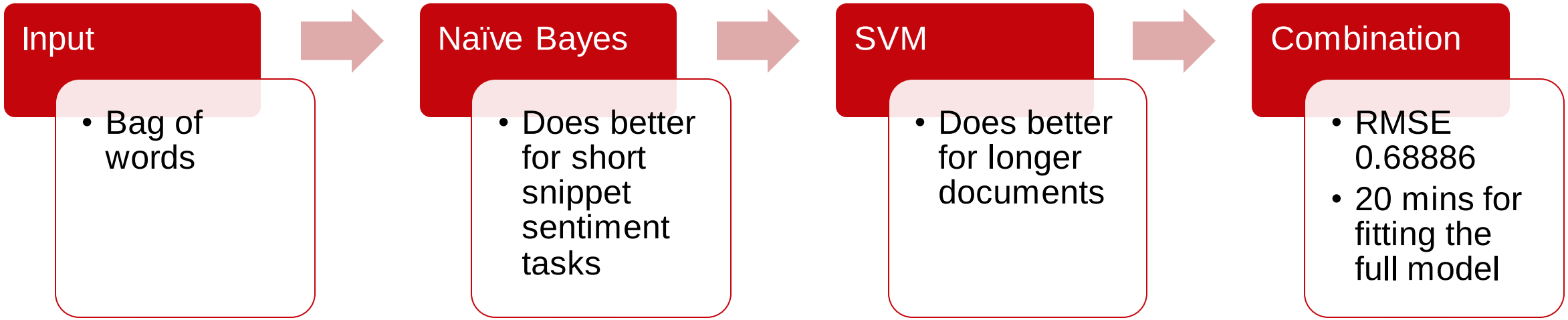
- Used to improve multiclass learning performance by separating multiple classes
- *"What is the probability that the review is a 1/2/3/4/5 star review?"*

	stars	1	2	3	4	5
6	1	True	False	False	False	False
7	4	False	False	False	True	False
8	3	False	False	True	False	False
9	3	False	False	True	False	False
10	5	False	False	False	False	True



Final Model

Naïve Bayes-Support Vector Machine(NBSVM)



XGBoost

Input

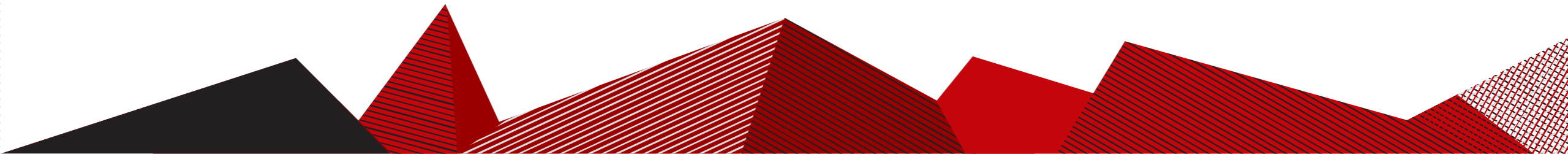
- Output of NBSVM
- Years
- Extra Features from reviews

Conclusion

- RMSE 0.58263
- 30 mins for fitting the full model

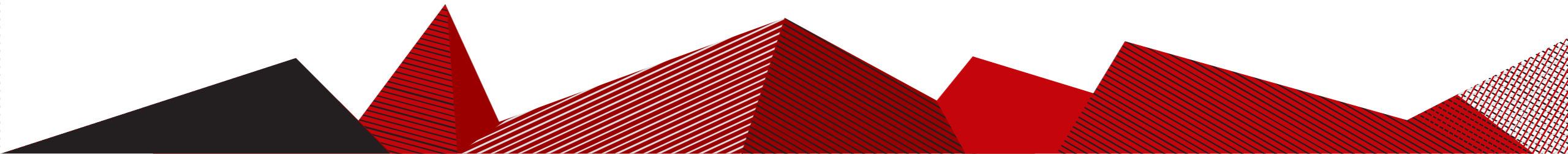
Reasons for combination

- Motivation of the stacking method
- Use XGBoost to extract extra information from the output of NBSVM



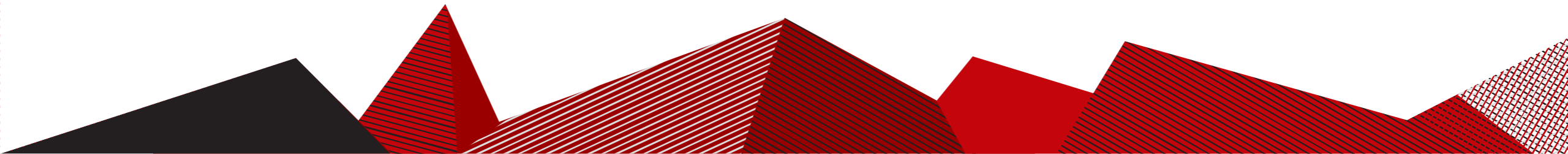
Strength

- Precise
- Fast
- Robust
- Simple



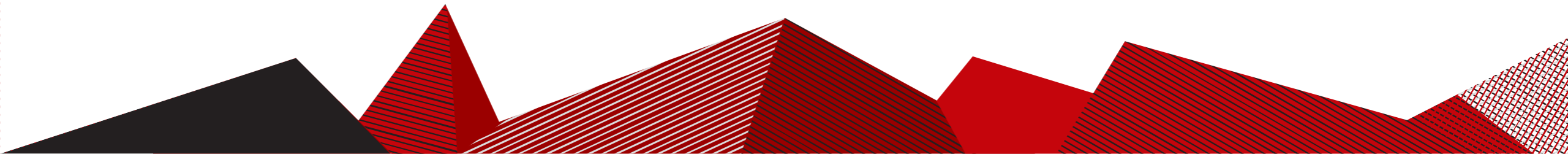
Weakness

- Typos
- Gap between online & offline RMSE



Conclusion

- Although we didn't gain the best score, our model is extremely **fast** to train and basically it's a **precise** and **reliable** model to predict sentiment levels of Yelp reviews.



Thank you!