ANALYSING AMAZON MOVIE REVIEWS

DATA SCIENCE IN TECHNO-SOCIO-ECONOMIC SYSTEMS

By Ferdinand Brunne, Oliver King, Brian Funk, and Silvan Metzker

MOTIVATION: REVIEW HELPFULNESS PROBLEM (RHP)

Understanding which Reviews are Helpful

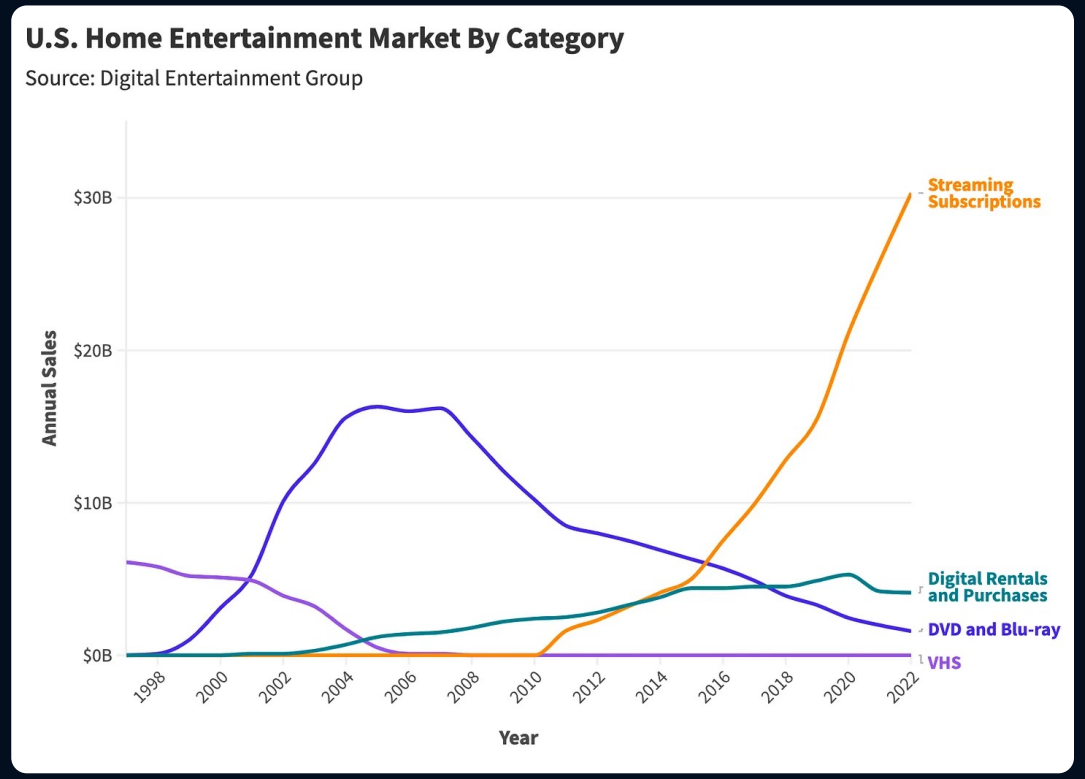
Display Positive Helpful Reviews

Display Negative Helpful Reviews

Helps Convince Consumer to buy Product

Alerts Customer of Valid Issues with Product

MOTIVATION: AMAZON MOVIE REVIEWS



Movie Sales across Formats in the U.S.A. [3]

Movies are of interest to the vast majority of people

A large amount of RHP research has already been done on movie reviews [1]

Largest Online Retailer: Word Record set in 2024, over $400 billion in annual sales [2]

Almost every movie in the world can be found on Amazon

Can we use a variety of Methods (ML, NN, and statistical models) to Predict the Helpfulness of Amazon Movie Reviews?

HYPOTHESIS

LITERATURE REVIEW

AN OVERVIEW

LITERATURE REVIEW

TIMELINE OF APPROACHES

Classical Machine Learning with SVM

"Automatically Assessing Review Helpfulness"

Sentiment Features

"Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics"

Transformer-Based Models (BERT)

"Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews"

MACHINE LEARNING WITH SVM 2006

AUTOMATICALLY ASSESSING REVIEW HELPFULNESS (KIM ET AL.) [5]

First to approach the challenge

SVM-based regression

Structural, lexical, syntactic, semantic and meta-data as features

Spearman correlation of 0.66

SENTIMENT FEATURES 2011

ESTIMATING THE HELPFULNESS AND ECONOMIC IMPACT OF PRODUCT REVIEWS: MINING TEXT AND REVIEWER CHARACTERISTICS (GHOSE ET AL.) [6]

Random Forst Classifiers for predicting helpfulness

Introduced new sentiment features for analysis: subjectivity & positivity

ATTENTION BASED MODELS 2022

EFFECTIVENESS OF FINE-TUNED BERT MODEL IN CLASSIFICATION OF HELPFUL AND UNHELPFUL ONLINE CUSTOMER REVIEWS (BILAL ET AL.) [7]

Using transformer architecture that only takes text as input

Part of larger shift towards using BERT for sentiment analysis

Showed that this approach could outperform pervious approaches

DATASET

OUR APPROACH

DATASET

By *Jure Leskovec* STANFORD UNIVERSITY

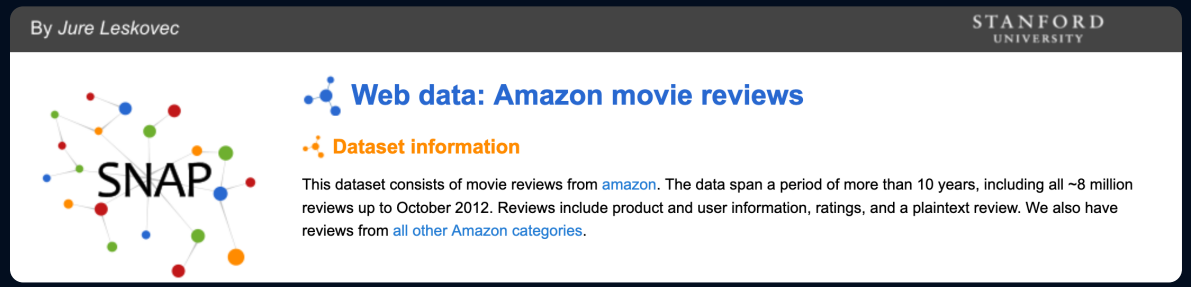
Web data: Amazon movie reviews

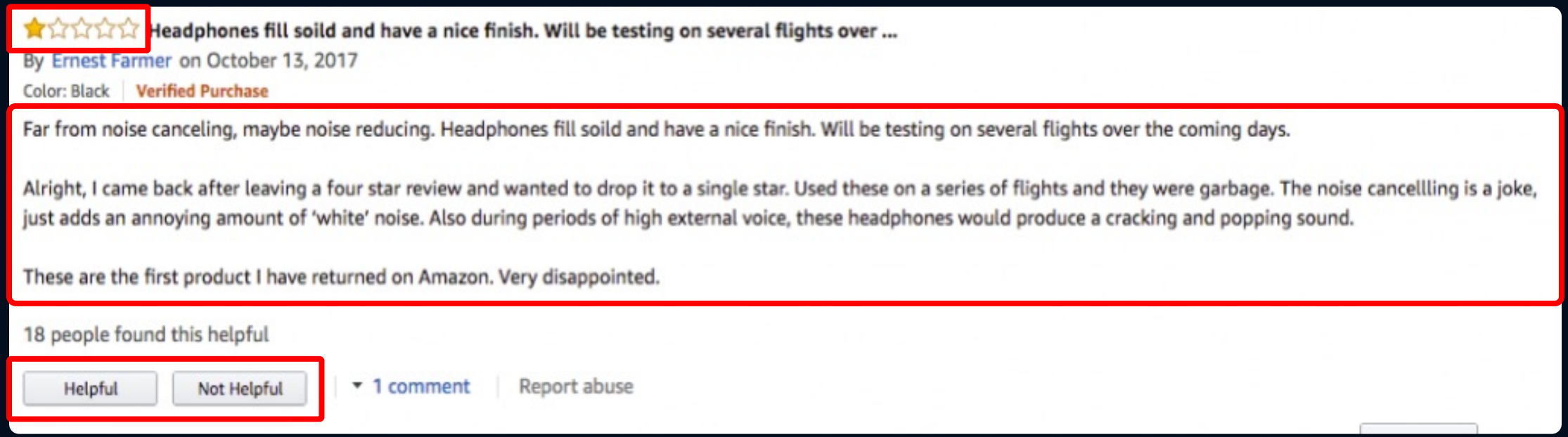
Dataset information

This dataset consists of movie reviews from amazon. The data span a period of more than 10 years, including all ~8 million reviews up to October 2012. Reviews include product and user information, ratings, and a plaintext review. We also have reviews from all other Amazon categories.

Web Data: Amazon Movie Reviews. Contains 7,911,684 total reviews. Collected from August 1997 to October 2012. [4]

DATASET

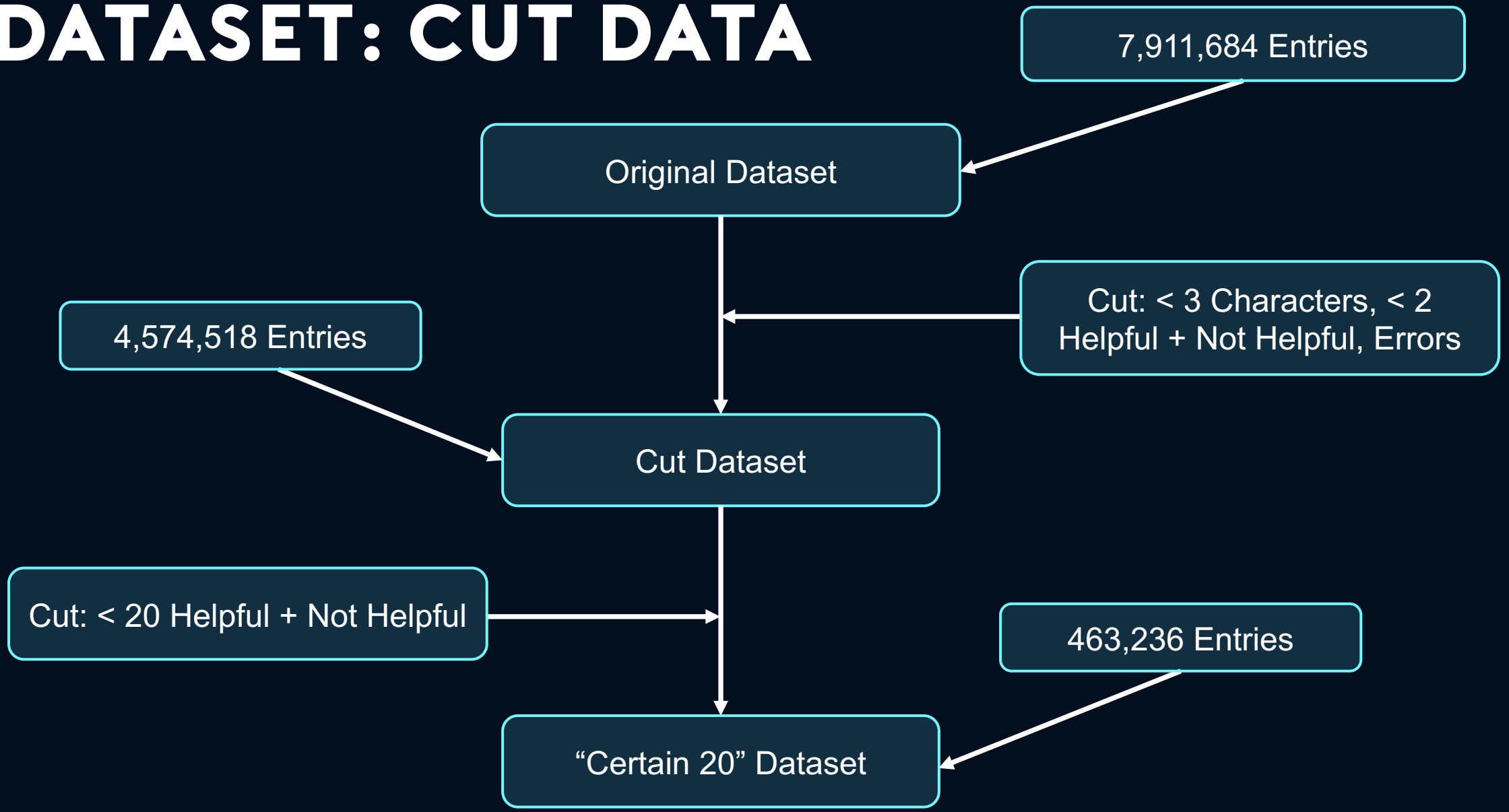

By Jure Leskovec
STANFORD UNIVERSITY
SNAP
Web data: Amazon movie reviews
Dataset information
This dataset consists of movie reviews from amazon. The data span a period of more than 10 years, including all ~8 million reviews up to October 2012. Reviews include product and user information, ratings, and a plaintext review. We also have reviews from all other Amazon categories.


Headphones fill soild and have a nice finish. Will be testing on several flights over ...
By Ernest Farmer on October 13, 2017
Color: Black
Verified Purchase
Far from noise canceling, maybe noise reducing. Headphones fill soild and have a nice finish. Will be testing on several flights over the coming days.
Alright, I came back after leaving a four star review and wanted to drop it to a single star. Used these on a series of flights and they were garbage. The noise cancellling is a joke, just adds an annoying amount of 'white' noise. Also during periods of high external voice, these headphones would produce a cracking and popping sound.
These are the first product I have returned on Amazon. Very disappointed.
18 people found this helpful
Helpful
Not Helpful
1 comment
Report abuse

```
product/productId: B00006HAXW
review/userId: A1RSDE90N6RSZF
review/profileName: Joseph M. Kotow
review/helpfulness: 9/9
review/score: 5.0
review/time: 1042502400
review/summary: Pittsburgh - Home of the OLDIES
review/text: I have all of the doo wop DVD's and this one is as good or better than the
1st ones. Remember once these performers are gone, we'll never get to see them again.
Rhino did an excellent job and if you like or love doo wop and Rock n Roll you'll LOVE
this DVD !!
```

Relevant to us: `Review/Helpfulness`, `Review/Score`, `Review/Summary`, `Review/Text`.

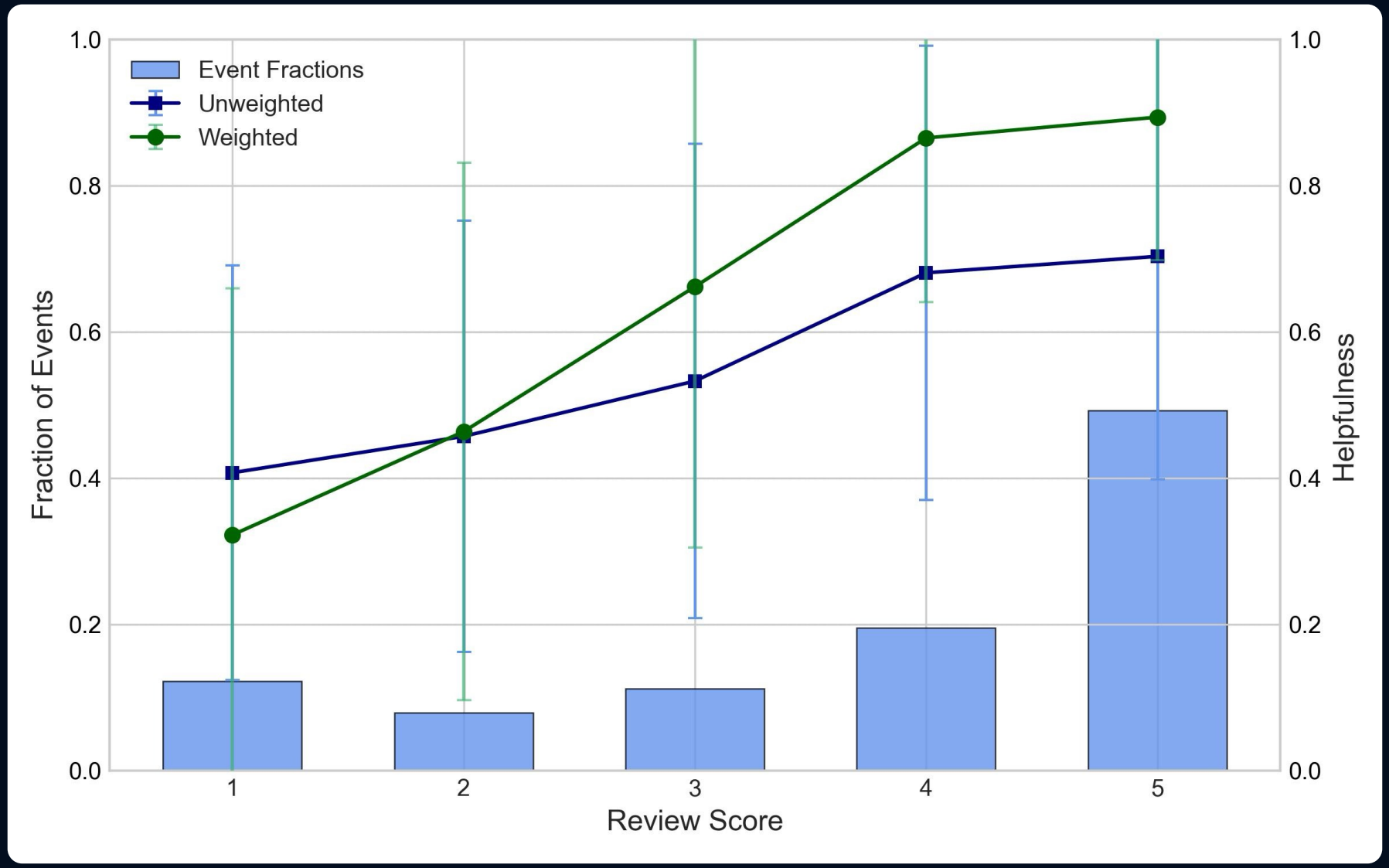
DATASET: CUT DATA



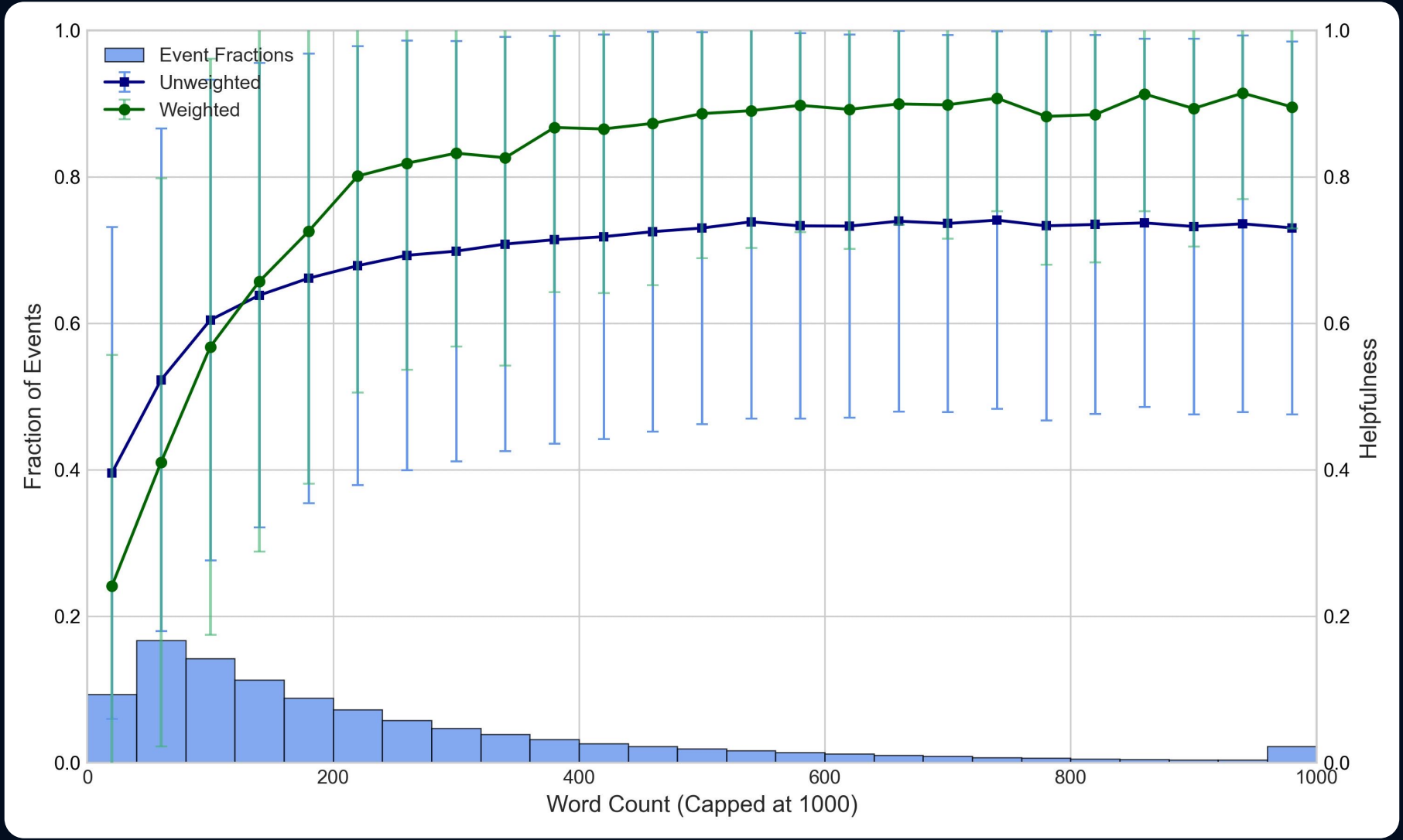
DATASET: EXPLANATORY DATA ANALYSIS (EDA)

	Rating	Word Count	Log(Word Count)	Helpfulness
Normality p	0	0	0	0
Spearman's	0.319655607	0.26753133	0.26753133	N/A
Spearman's p	0	0	0	N/A
Average	3.855191957	239.8625407	5.048032972	0.623933606
Std. Dev.	1.414071519	249.3789082	0.95199141	0.324989332
Weighted Av.	3.796418505	379.8200473	5.535315764	0.743331321
W Std. Dev.	1.554992361	368.8492116	0.958509263	0.342620396

DATASET: EDA: RATING



DATASET: EDA: WORD COUNT



DATA AUGMENTATION

OUR APPROACH

VADER

METHOD

“terrible” = -2.1

“great” = 3.1

WORD SCORING

”not great” = -3.1

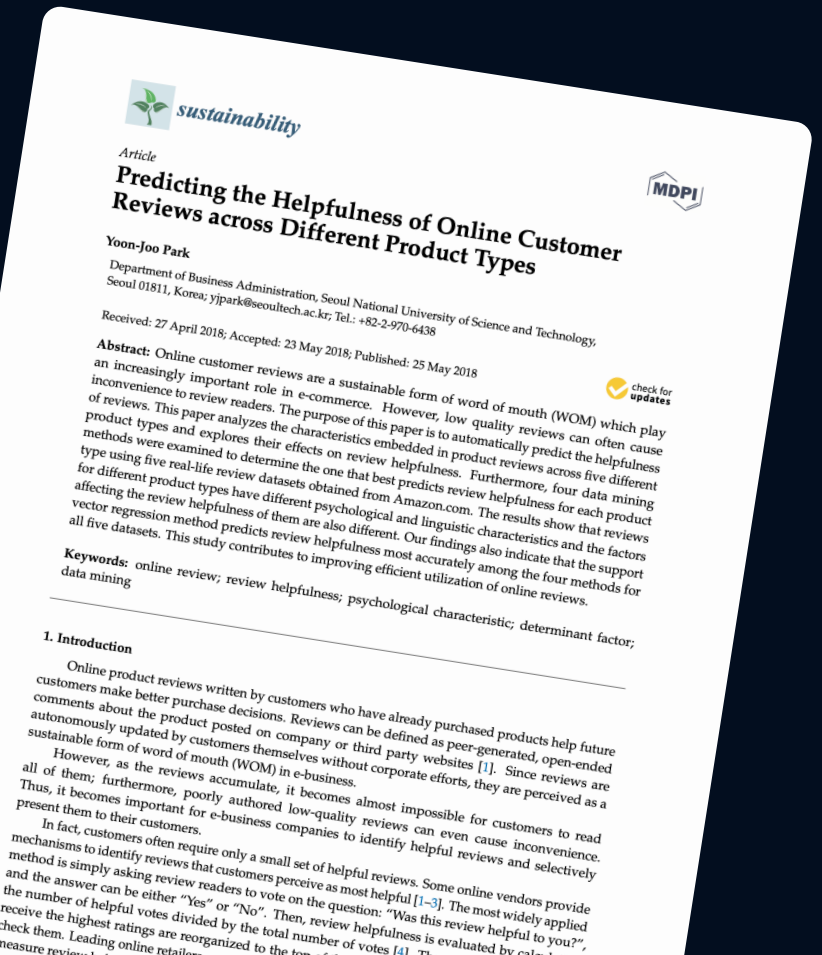
NEGATIONS

“!” or “very”

DEGREE MODIFIERS

$$\texttt{sentiment\_score} \in [-1, 1]$$

EMPATH
METHOD



sustainability

Article

Predicting the Helpfulness of Online Customer Reviews across Different Product Types

Yoon-Joo Park

Department of Business Administration, Seoul National University of Science and Technology, Seoul 01811, Korea; yjpark@seoultech.ac.kr; Tel.: +82-2-970-6438

Received: 27 April 2018; Accepted: 23 May 2018; Published: 25 May 2018

Abstract: Online customer reviews are a sustainable form of word of mouth (WOM) which play an increasingly important role in e-commerce. However, low quality reviews can often cause inconvenience to review readers. The purpose of this paper is to automatically predict the helpfulness of reviews. This paper analyzes the characteristics embedded in product reviews across five different product types and explores their effects on review helpfulness. Furthermore, four data mining methods were examined to determine the one that best predicts review helpfulness for each product type using five real-life review datasets obtained from Amazon.com. The results show that reviews for different product types have different psychological and linguistic characteristics and the factors affecting the review helpfulness of them are also different. Our findings also indicate that the support vector regression method predicts review helpfulness most accurately among the four methods for all five datasets. This study contributes to improving efficient utilization of online reviews.

Keywords: online review; review helpfulness; psychological characteristic; determinant factor; data mining

1. Introduction

Online product reviews written by customers who have already purchased products help future customers make better purchase decisions. Reviews can be defined as peer-generated, open-ended comments about the product posted on company or third party websites [1]. Since reviews are autonomously updated by customers themselves without corporate efforts, they are perceived as a sustainable form of word of mouth (WOM) in e-business.

However, as the reviews accumulate, it becomes almost impossible for customers to read all of them; furthermore, poorly authored low-quality reviews can even cause inconvenience. Thus, it becomes important for e-business companies to identify helpful reviews and selectively present them to their customers.

In fact, customers often require only a small set of helpful reviews. Some online vendors provide mechanisms to identify reviews that customers perceive as most helpful [1–3]. The most widely applied method is simply asking review readers to vote on the question: "Was this review helpful to you?", and the answer can be either "Yes" or "No". Then, review helpfulness is evaluated by calculating the number of helpful votes divided by the total number of votes [4]. The highest ratings are reorganized to the top, receive the highest ratings are reorganized to the top, check them. Leading online retailers measure review

Table 4. Explanation of the research variables.

Variable	Explanation	Calculation
Rating	Rating score of a product from a reviewer scaled from 1 to 5	Rating score of a product
WC	Total number of words included in the review text	Word count
WPS	Average number of words in a sentence	# of words/# of sentences
Compare	Ratio of the number of comparison words (bigger, best, smaller, etc.) in the review text to a total of 317 comparison words in the LIWC 2015 dictionary	(# of related words in the review text/total # of related words) × 100
Analytic	Level of formal, logical, and hierarchical thinking scaled from 0 to 100. Lower numbers reflect more informal, personal, here and now, and narrative thinking.	Derived based on previously published findings from Pennebarker et al. [30]
Clout	Level of expertise and confident thinking scaled from 0 to 100. Low Clout numbers suggest a more tentative, humble, and even anxious style.	Derived based on previously published findings from Kacewicz et al. [31]
Authentic	Level of honest, personal, and disclosing thinking scaled from 0 to 100. Lower numbers suggest a more guarded, distanced form of discourse.	Derived based on previously published findings from Newman et al. [32]
CogProc	Ratio of the number of cognitive process words (cause, know, ought, etc.) in the review text to a total of 797 cognitive words in the LIWC 2015 dictionary	(# of related words in a review text/total # of related words) × 100
Percept	Ratio of the number of perceptual process words (look, heard, feeling, etc.) in the review text to a total of 436 perceptual words in the LIWC 2015 dictionary	(# of related words in a review text/total # of related words) × 100
PosEmo	Ratio of the number of positive emotion words (love, nice, sweet, etc.) in the review text to a total of 620 negative emotion words in the LIWC 2015 dictionary	(# of related words in a review text/total # of related words) × 100
NegEmo	Ratio of the number of negative emotion words (hurt, ugly, nasty, etc.) in the review text to a total of 744 negative emotion words in the LIWC 2015 dictionary	(# of related words in a review text/total # of related words) × 100
Helpfulness	Ratio of the number of helpful votes to the total number of votes	(Helpful #/Total #) × 100

https://www.mdpi.com/2071-1050/10/6/1735

OUR METRICS

METHOD

Rating: Product rating score received from a reviewer.

WC: The length of a review measured by the number of words in the review text (Word Count).

Analytic: The level of formal, logical, and hierarchical thinking.

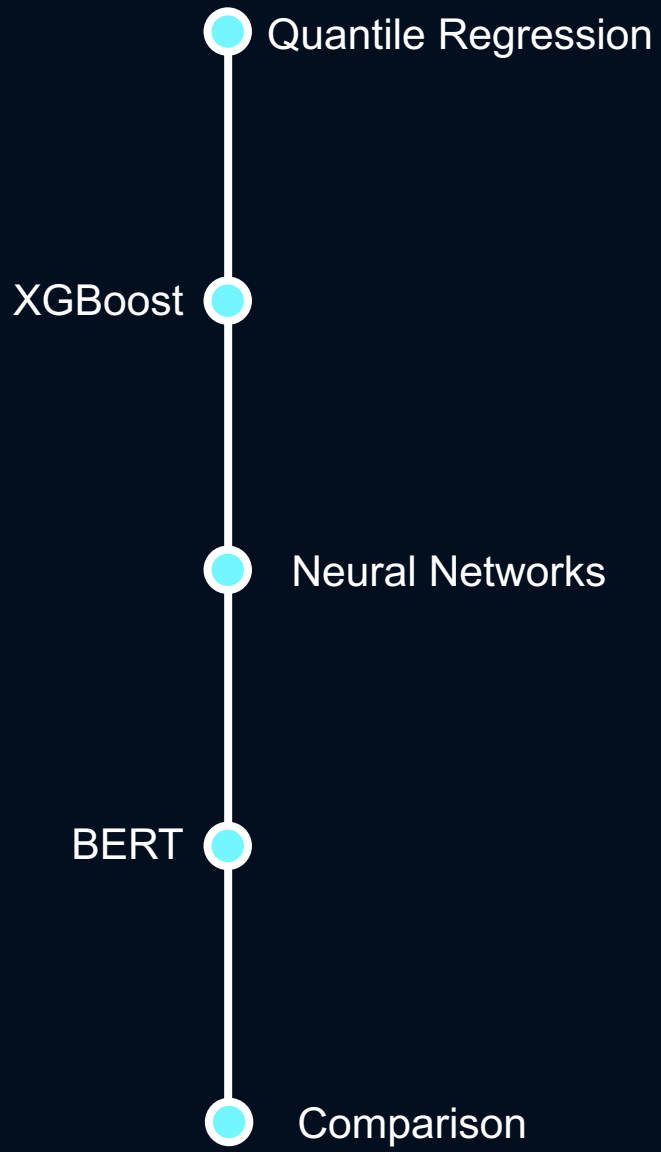
Sentiment Score: VADER Score estimating sentiment [-1,1].

FleschKincaid Score: Readability metric that indicates how easy a text is to understand.

https://www.mdpi.com/2071-1050/10/6/1735

OUR MODELS

OUR APPROACH


Quantile Regression
XGBoost
Neural Networks
BERT
Comparison

WEIGHTINGS FOR STATISTICAL MODELS

$$H = \frac{H^+}{H^+ + H^-} \xrightarrow{\text{Laplace Smoothing}} H = \frac{H^+ + 1}{H^+ + H^- + 2}$$

HELPFULNESS IN STANDARD ERROR

$$\mathrm{SE} = \sqrt{\frac{H(1-H)}{n}}$$

STANDARD ERROR

$$w = \frac{1}{\mathrm{SE}^2}$$

WEIGHT

ERROR MEASURE

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |H_i - \widehat{H}_i|}{n}$$

MEAN ABSOLUTE ERROR

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (H_i - \widehat{H}_i)^2}{n}$$

MEAN SQUARED ERROR

$$\mathrm{WMAE} = \frac{\sum_{i=1}^{n} w_i |H_i - \widehat{H}_i|}{\sum_{i=1}^{n} w_i}$$

WEIGHTED MEAN ABSOLUTE ERROR

$$\mathrm{WMSE} = \frac{\sum_{i=1}^{n} w_i (H_i - \widehat{H}_i)^2}{\sum_{i=1}^{n} w_i}$$

WEIGHTED MEAN SQUARE ERROR

QUANTILE REGRESSION

SETUP

Quantile Regression Model **run ten times**, results are average and standard deviation of these runs

Quantile Regression intentionally did not have its loss completely minimised, to **avoid overfitting** at edges (explained later)

Took a **Weighted Approach**

Each input parameter was **mean-centred**, except Word Count which was logged, and a flexible sigmoid was applied to the results for better fitting

$$H = \frac{1}{1 + e^{-\frac{H'}{1.25}}}$$

FLEXIBLE SIGMOID

QUANTILE REGRESSION

METHOD

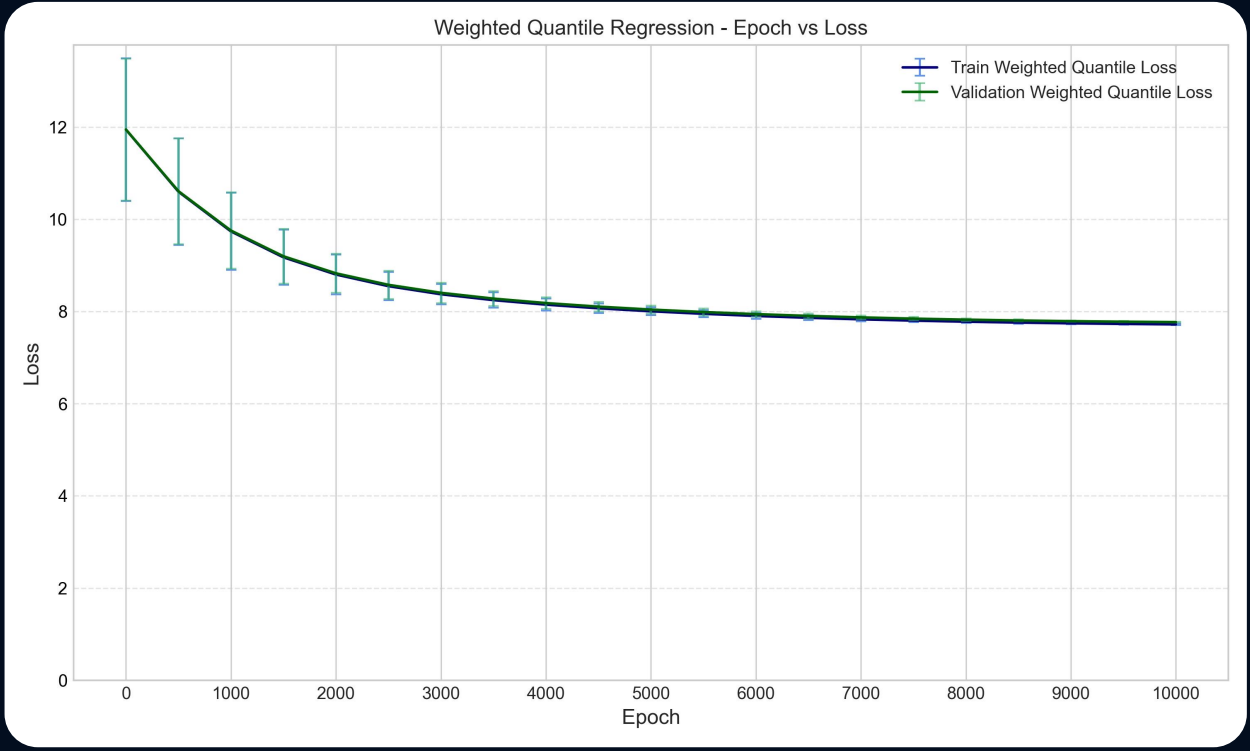
Estimates the conditional quantile: for a quantile of 0.5, this is the median

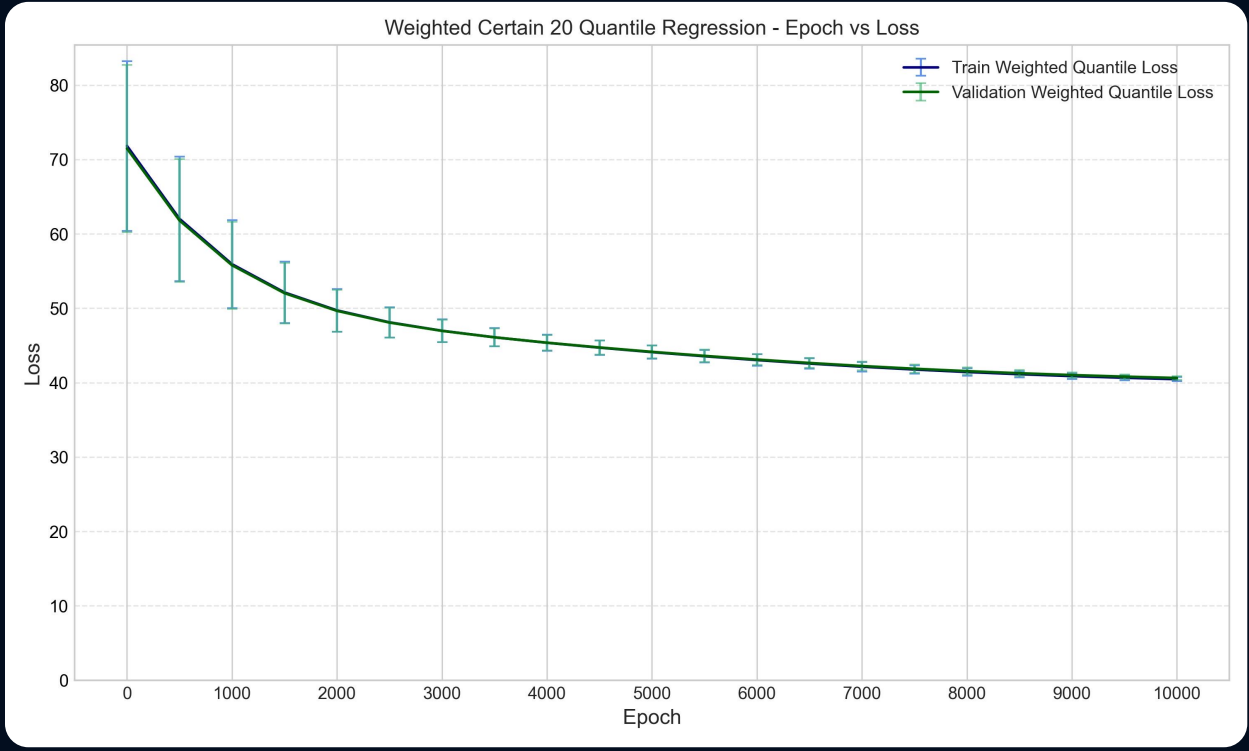
Minimises Weighted Quantile Loss: we chose a quantile of 0.5, it becomes exactly half the Weighted MAE

Does **not** rely on assumption of linearity, unlike linear regression

QUANTILE REGRESSION

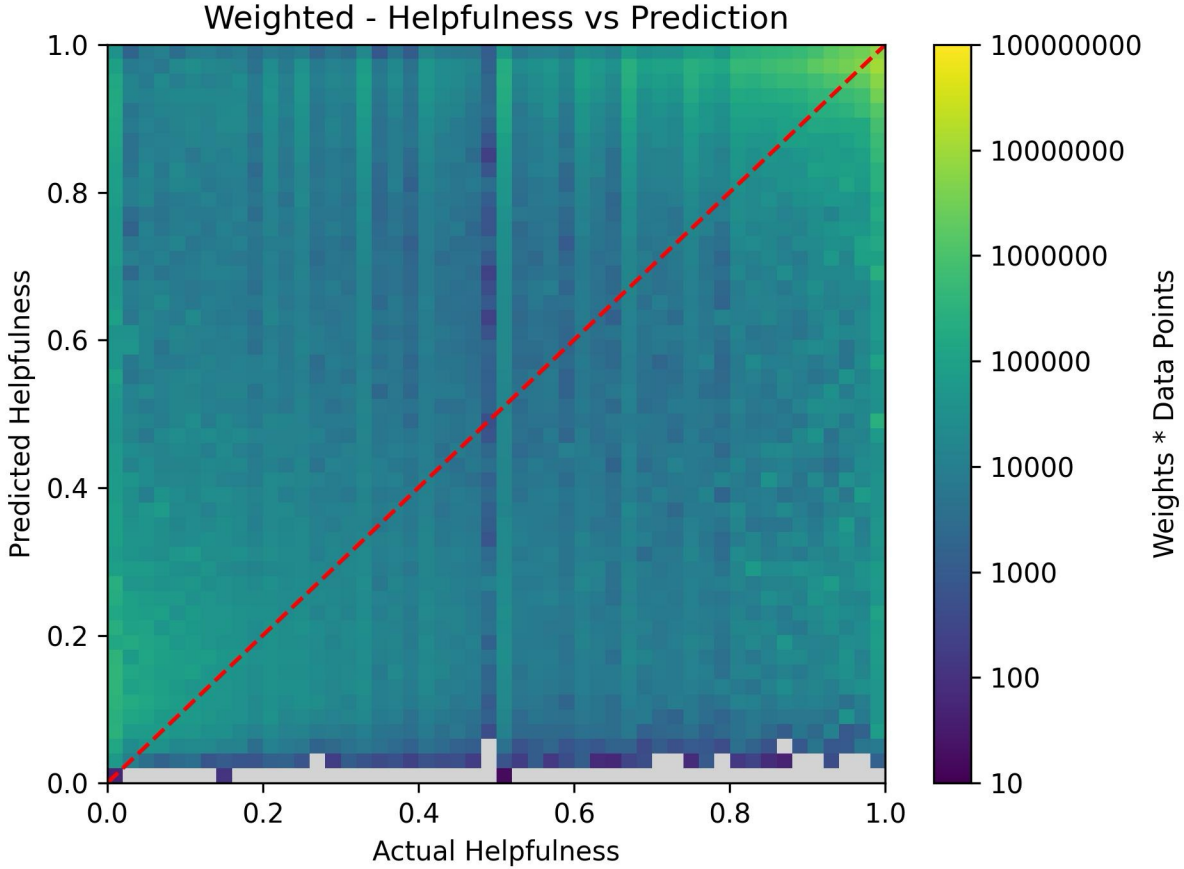
RESULTS

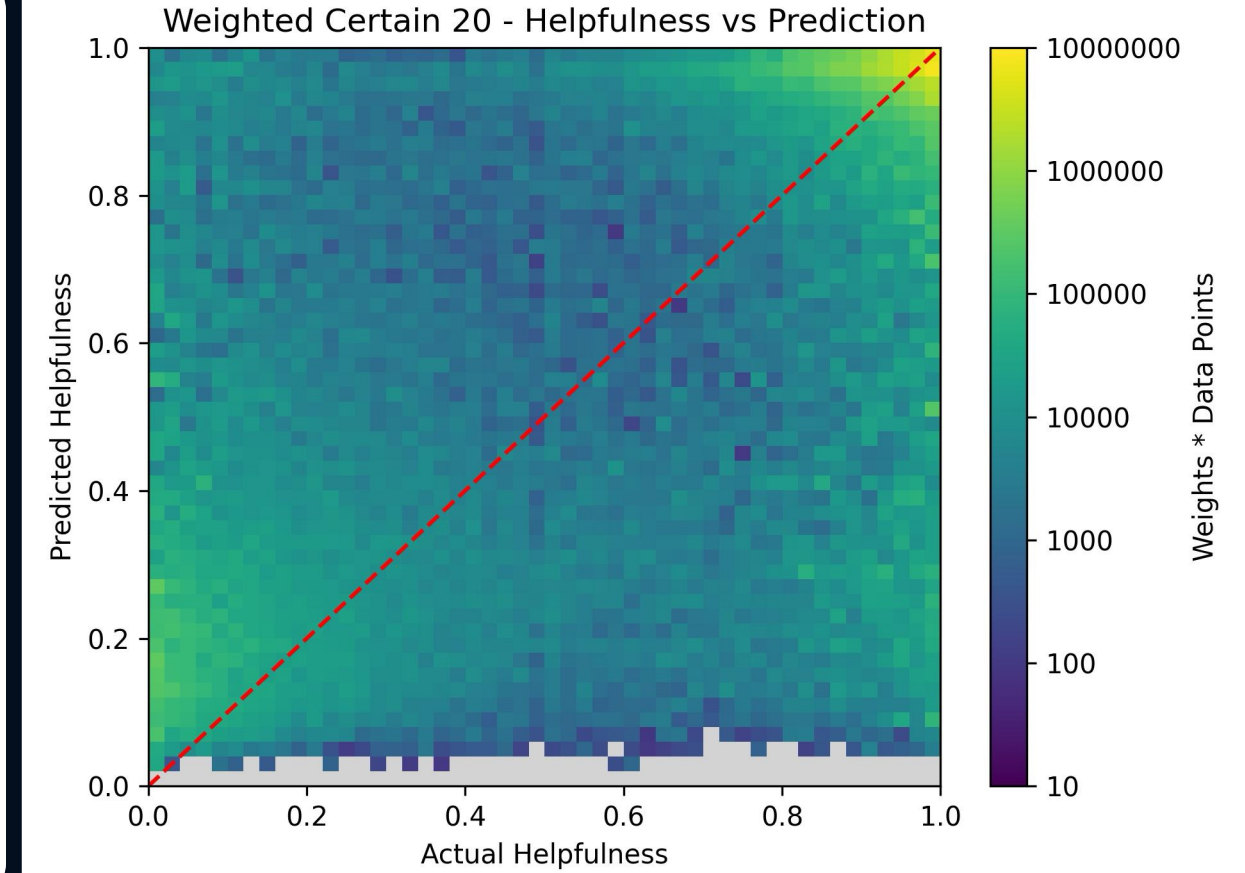




QUANTILE REGRESSION

RESULTS





QUANTILE REGRESSION

RESULTS

* : These values are Mean-Centred

$$y' = \text{Helpfulness (Pre}-\text{Sigmoid)}$$
$$x_1 = \text{Log(Word Count)}$$
$$x_2 = \text{Sentiment Score}^*$$
$$x_3 = \text{User Score}^*$$
$$x_4 = \text{Analytic Score}^*$$
$$x_5 = \text{Readability Score}^*$$
$$c = 1$$

$$y' = (0.462 \pm 0.018)x_1 + (1.38 \pm 0.05)x_2 + (1.26 \pm 0.04)x_3 + (0.62 \pm 0.04)x_4 + (0.156 \pm 0.004)x_5 - (0.08 \pm 0.08)c$$

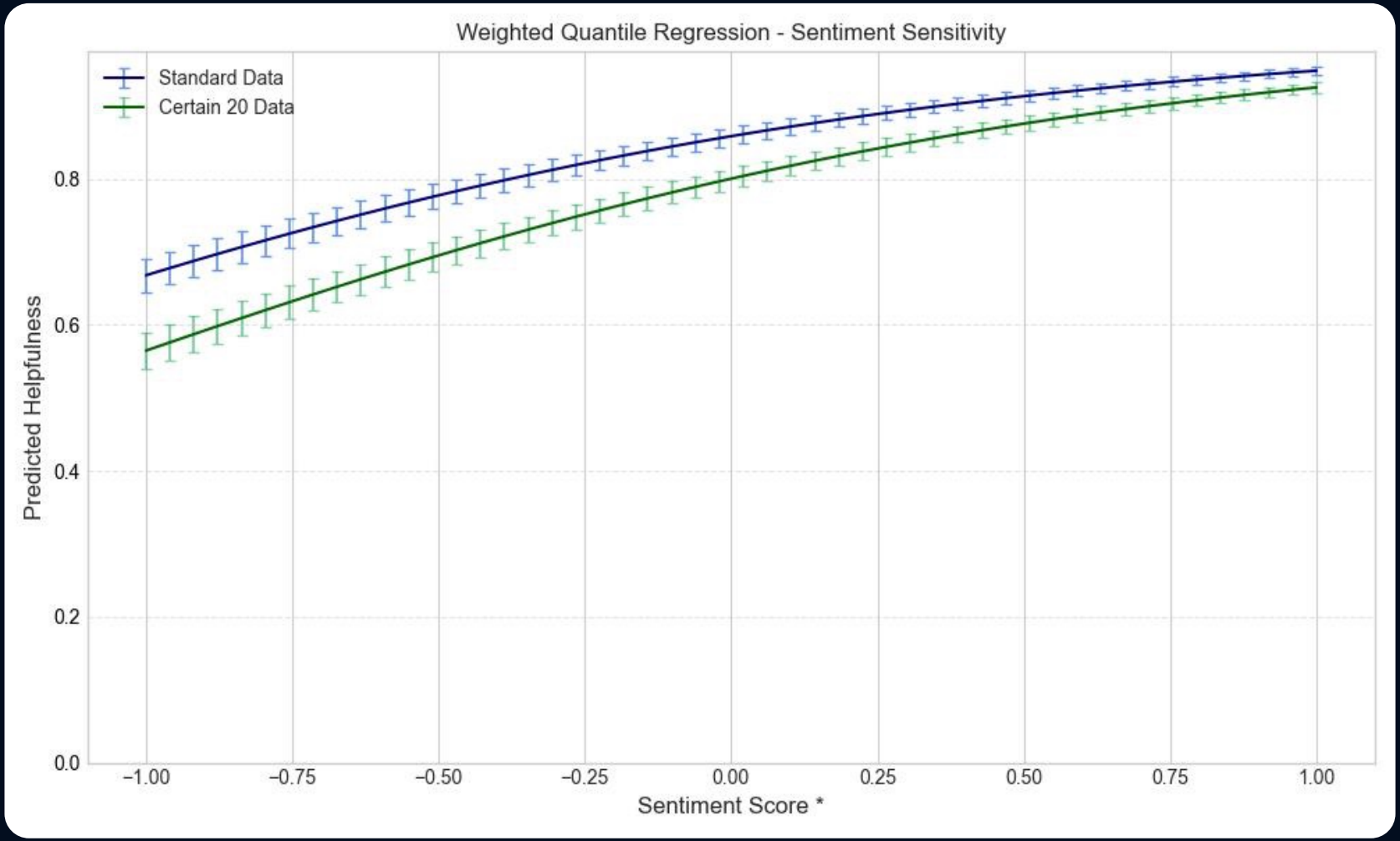
Full Dataset

$$y' = (0.360 \pm 0.016)x_1 + (1.41 \pm 0.06)x_2 + (1.30 \pm 0.06)x_3 + (0.53 \pm 0.06)x_4 + (0.162 \pm 0.008)x_5 - (0.22 \pm 0.07)c$$

Certain 20

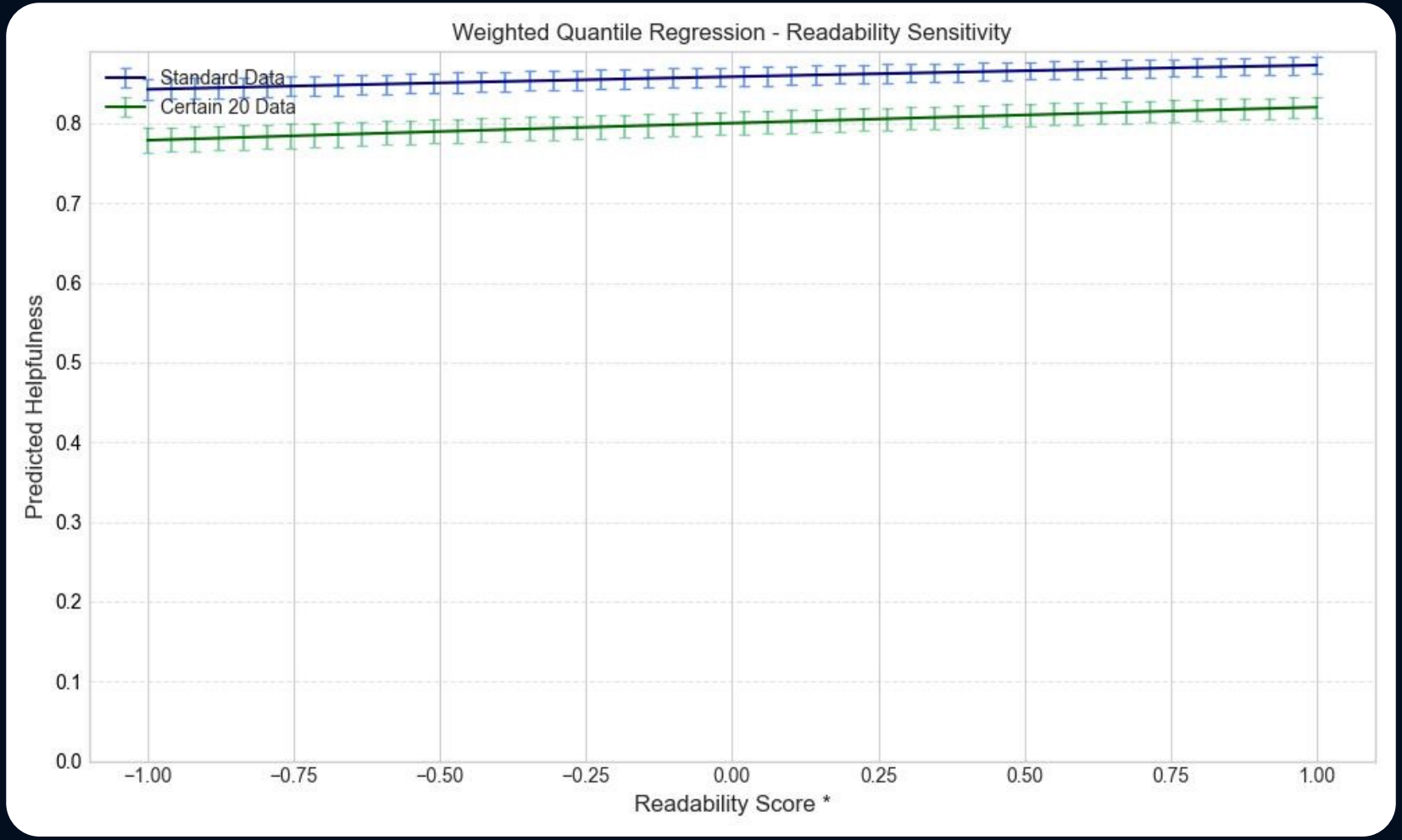
QUANTILE REGRESSION

SENSITIVITY ANALYSIS



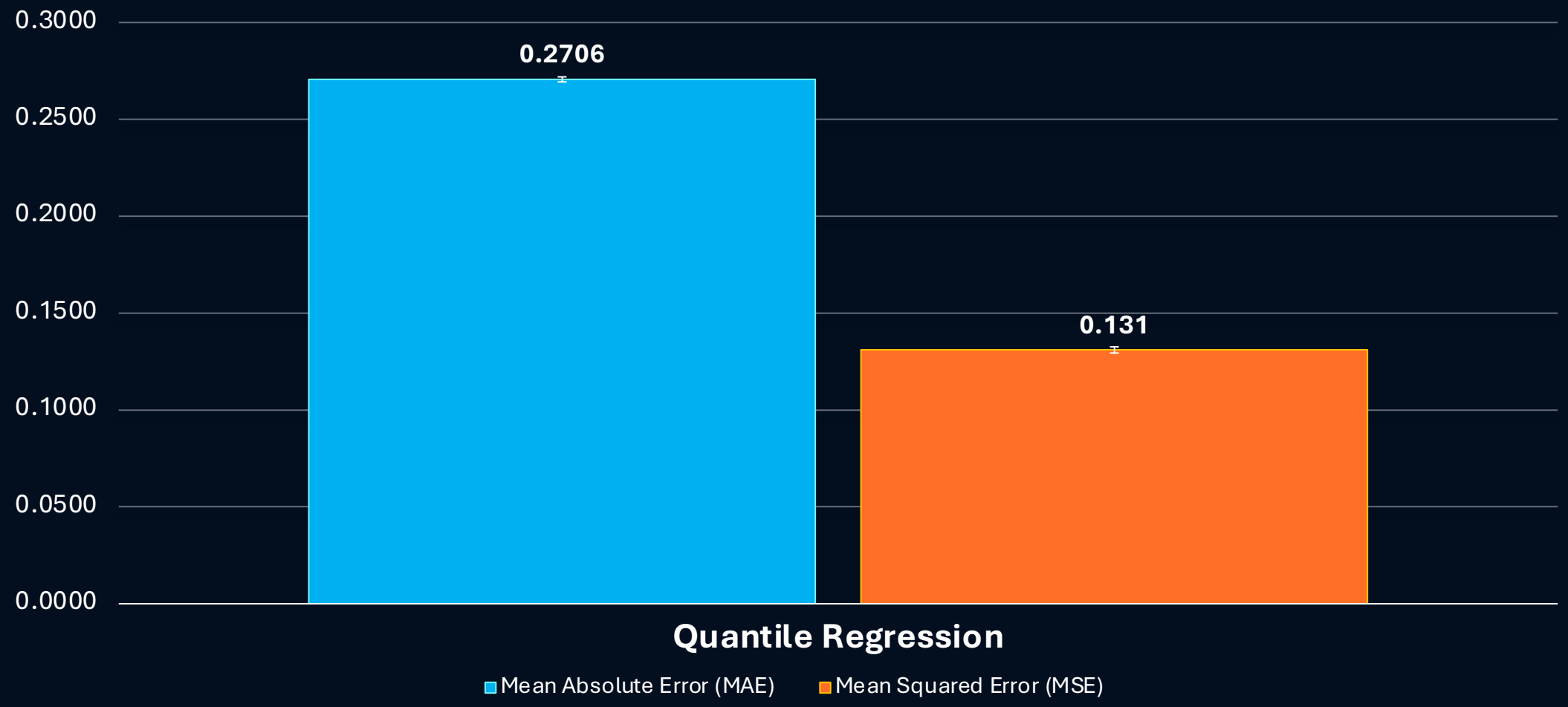
QUANTILE REGRESSION

SENSITIVITY ANALYSIS



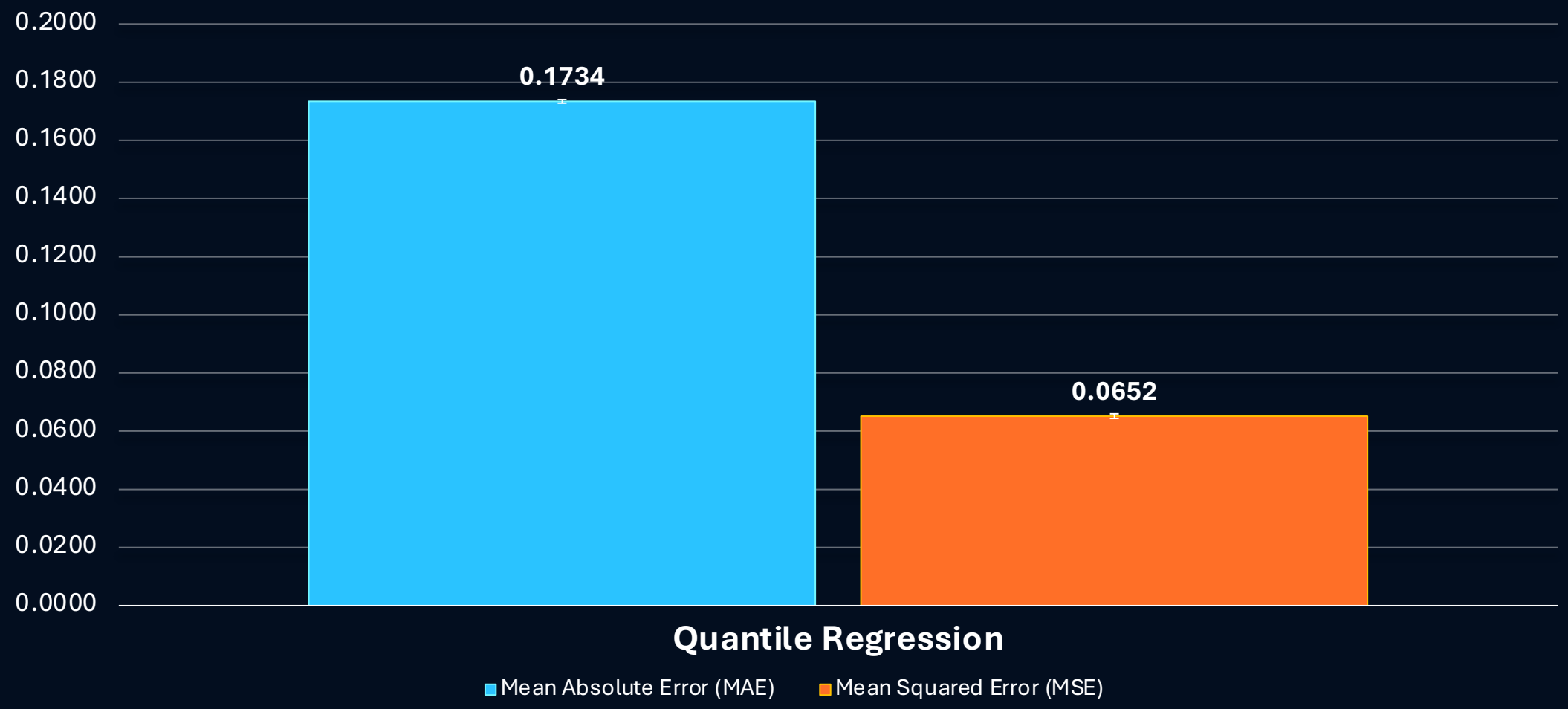
FULL DATASET

RESULTS



CERTAIN TWENTY

RESULTS



XGBOOST

METHOD

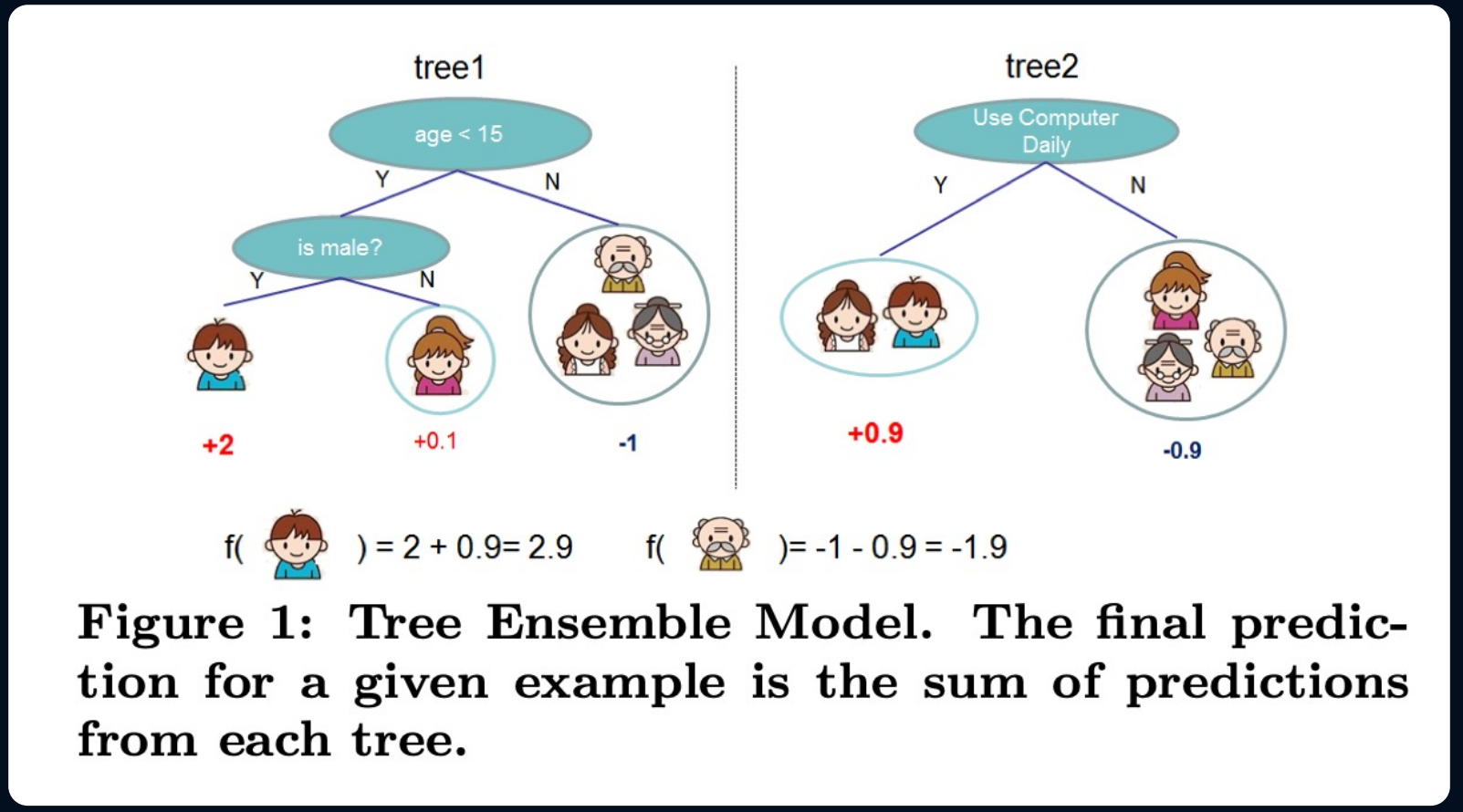
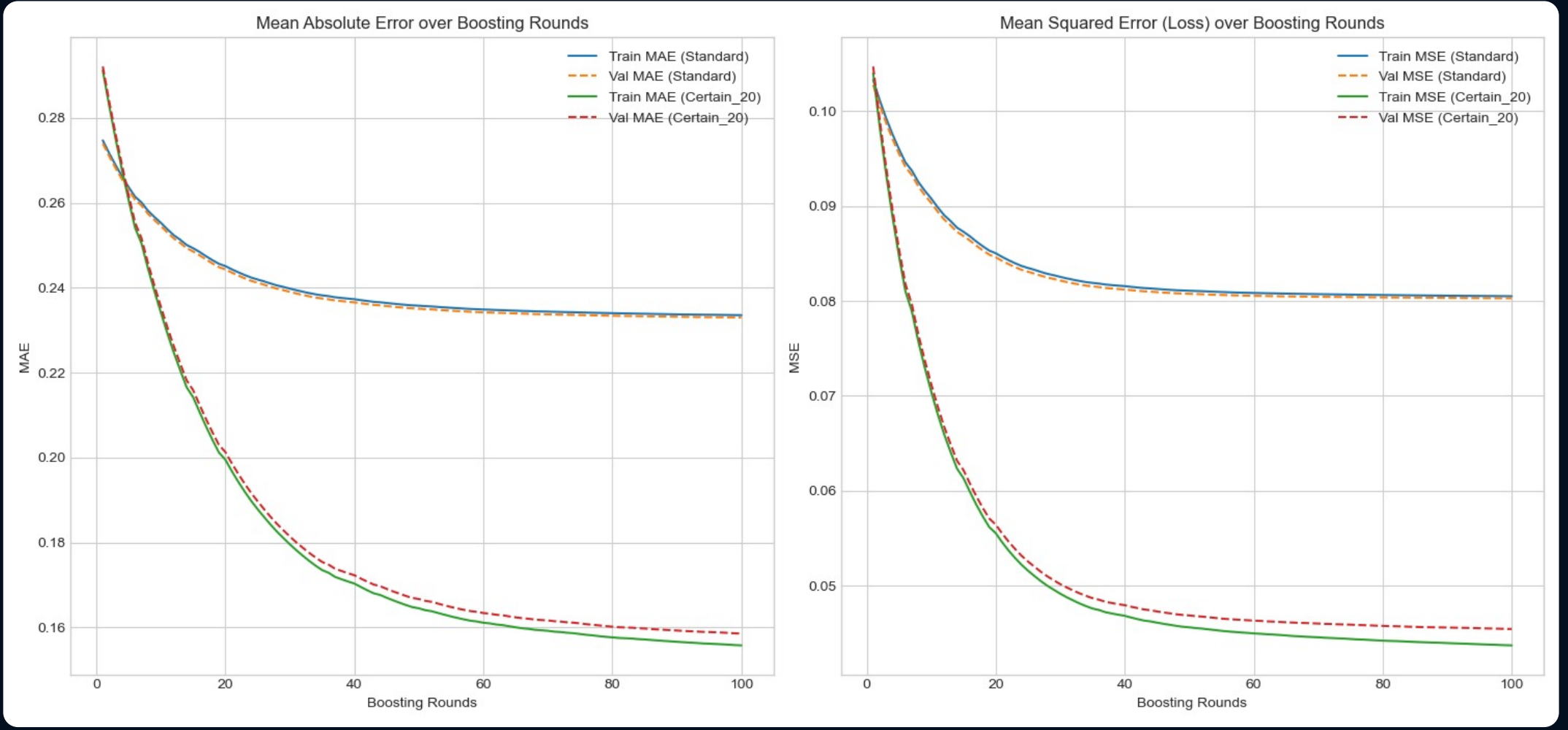


Figure 1: Tree Ensemble Model. The final prediction for a given example is the sum of predictions from each tree.

https://arxiv.org/abs/1603.02754 [8]

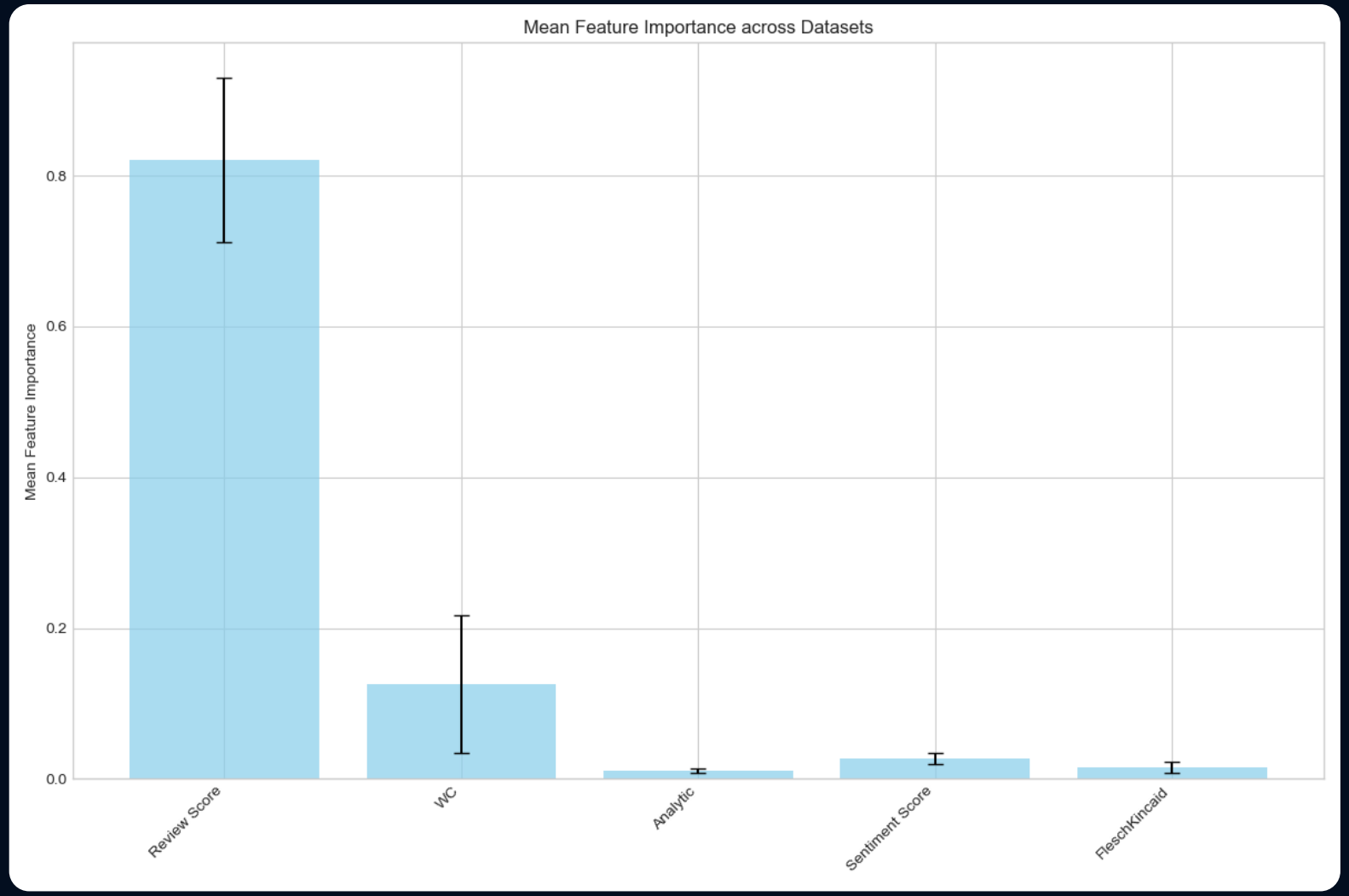
XGBOOST

RESULTS



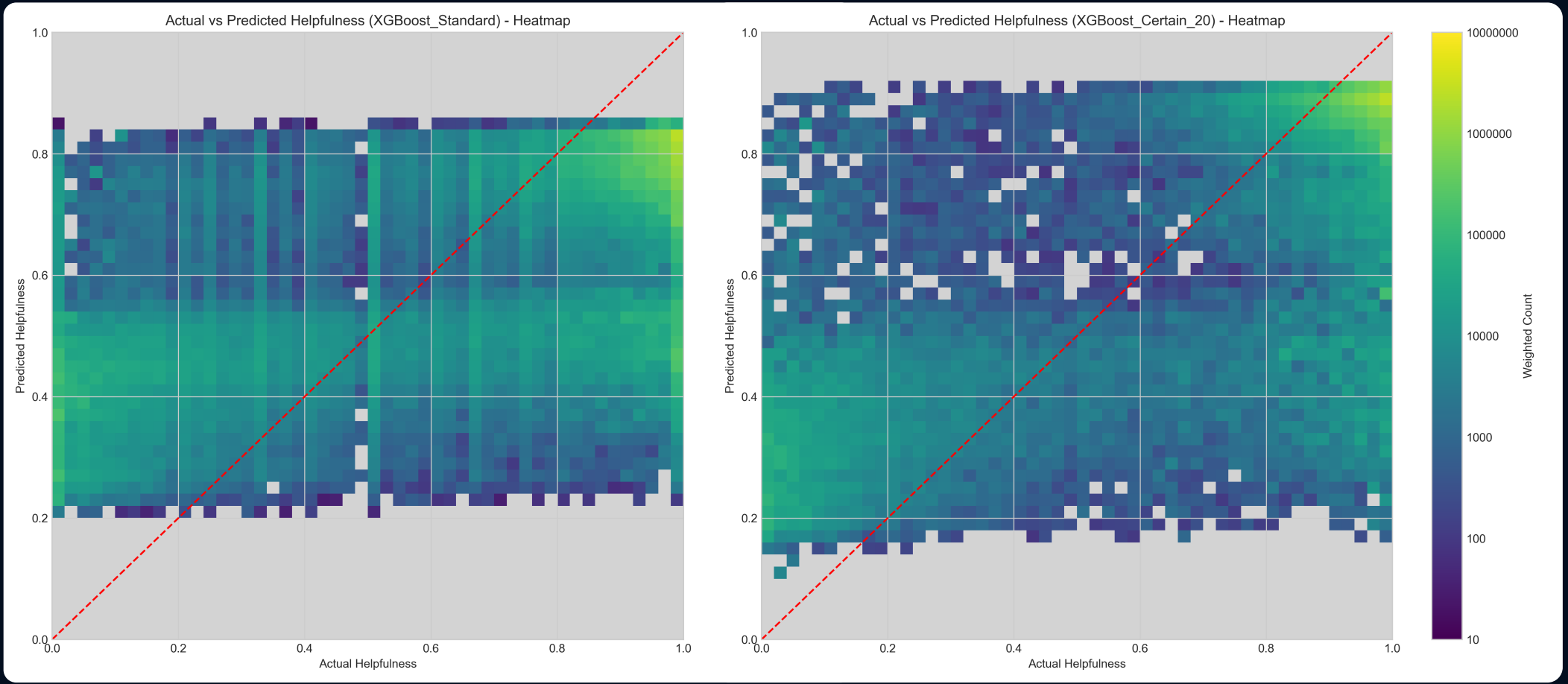
XGBOOST

RESULTS



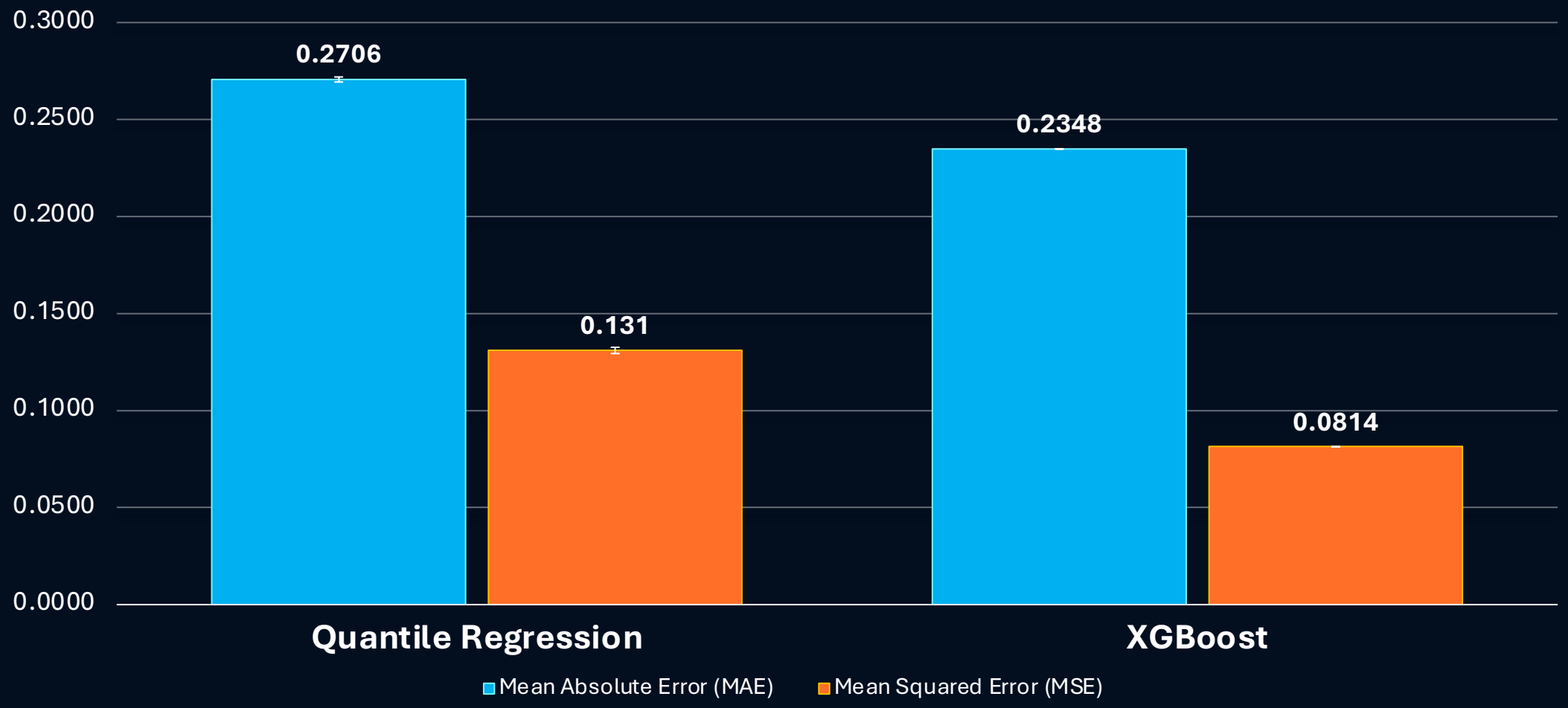
XGBOOST

RESULTS



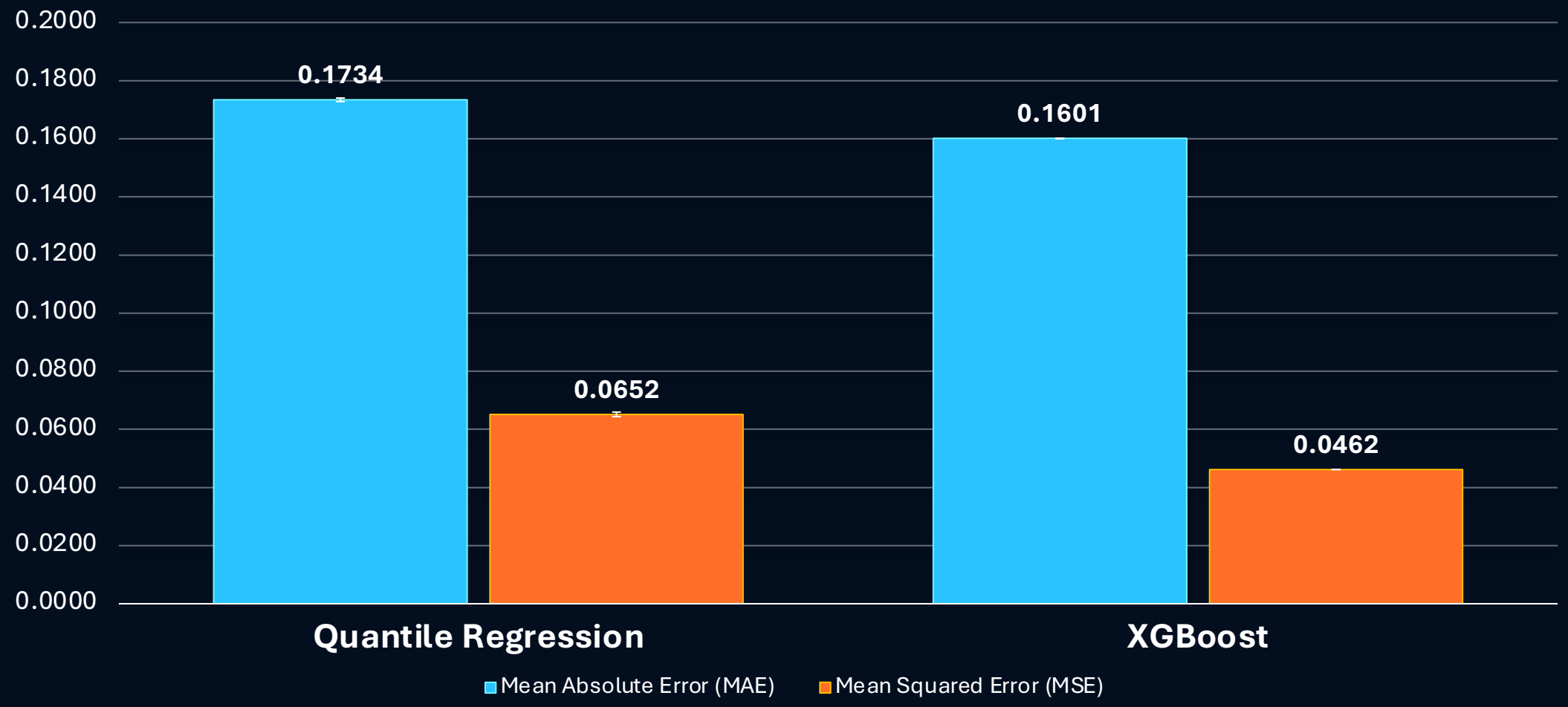
FULL DATASET

RESULTS



CERTAIN TWENTY

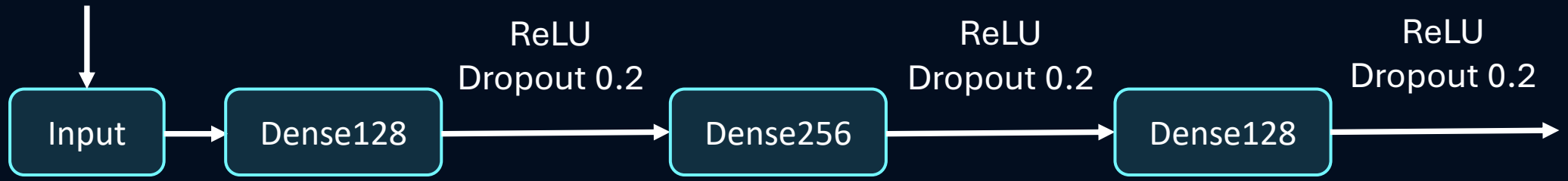
RESULTS



NEURAL NETWORK W/ EMPATH

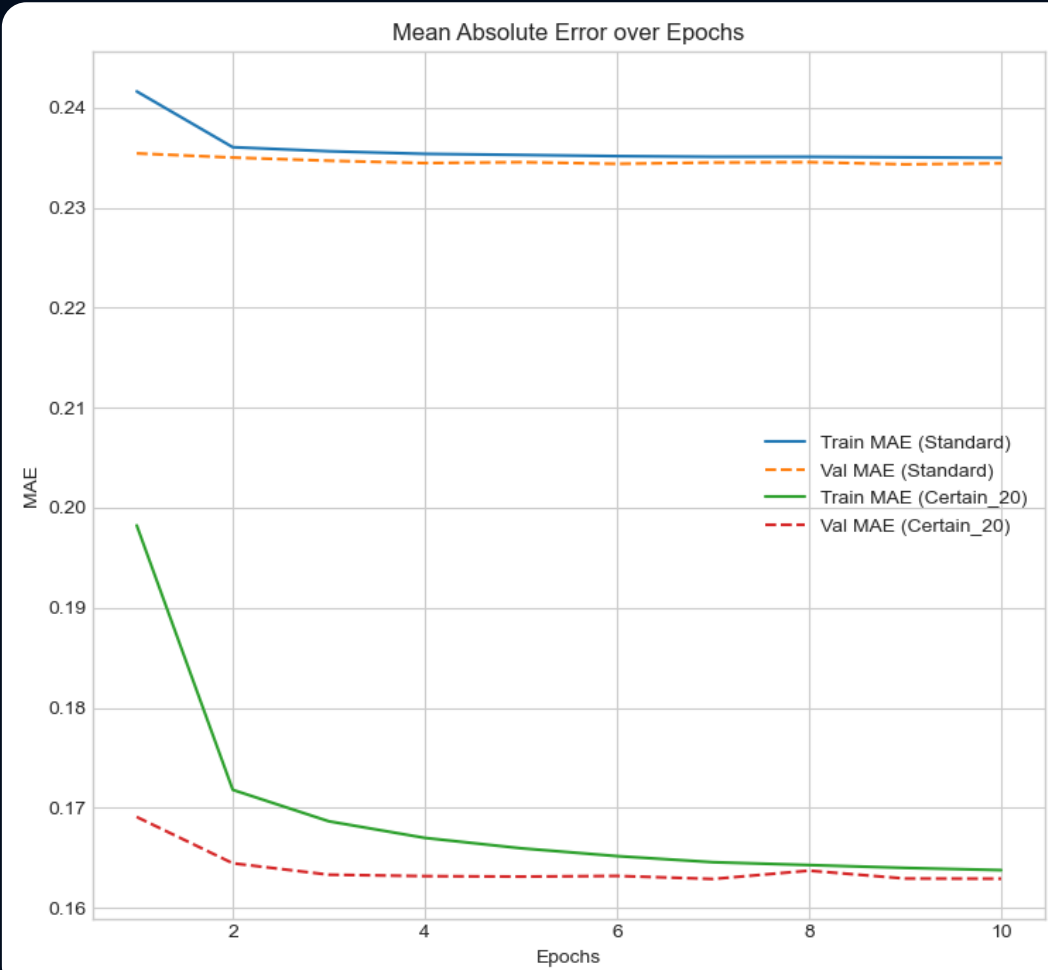
METHOD

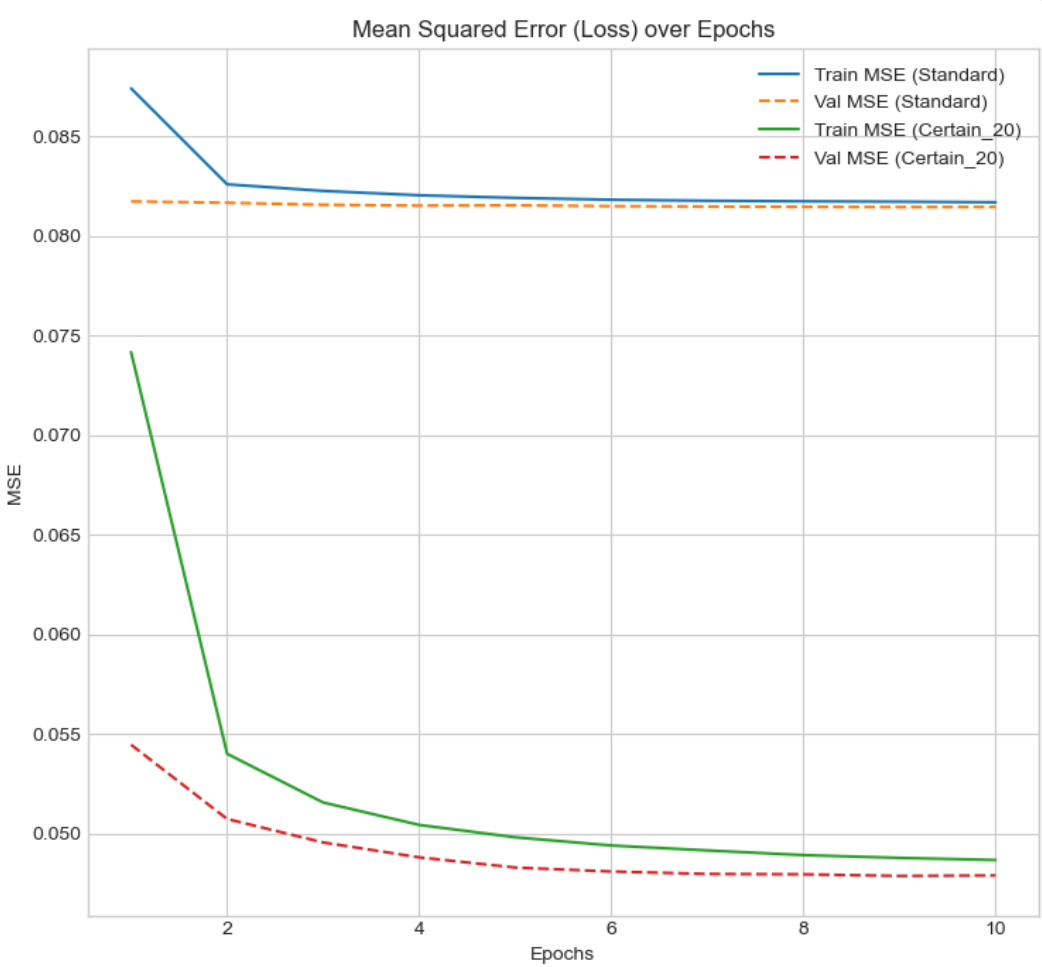
```
review/length
review/score
sentiment_score
```



NEURAL NETWORK W/ EMPATH

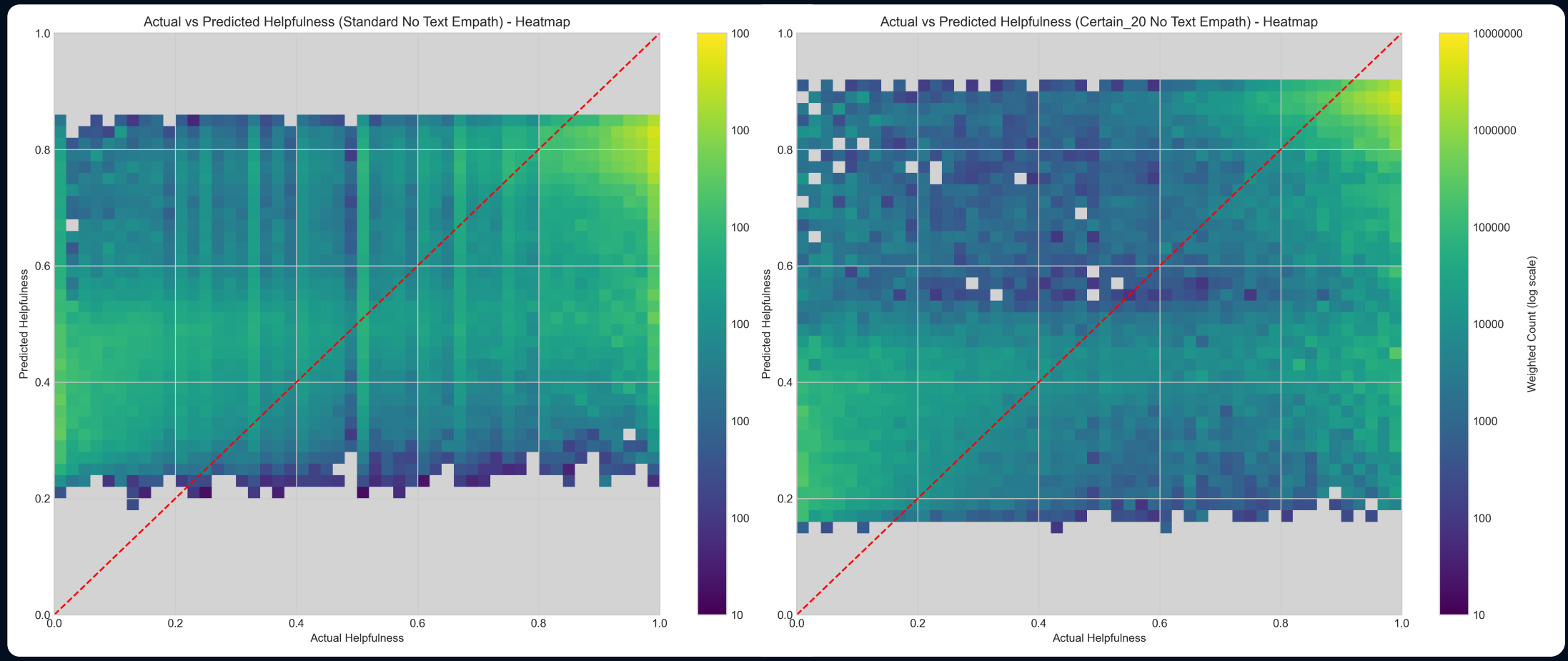
RESULTS





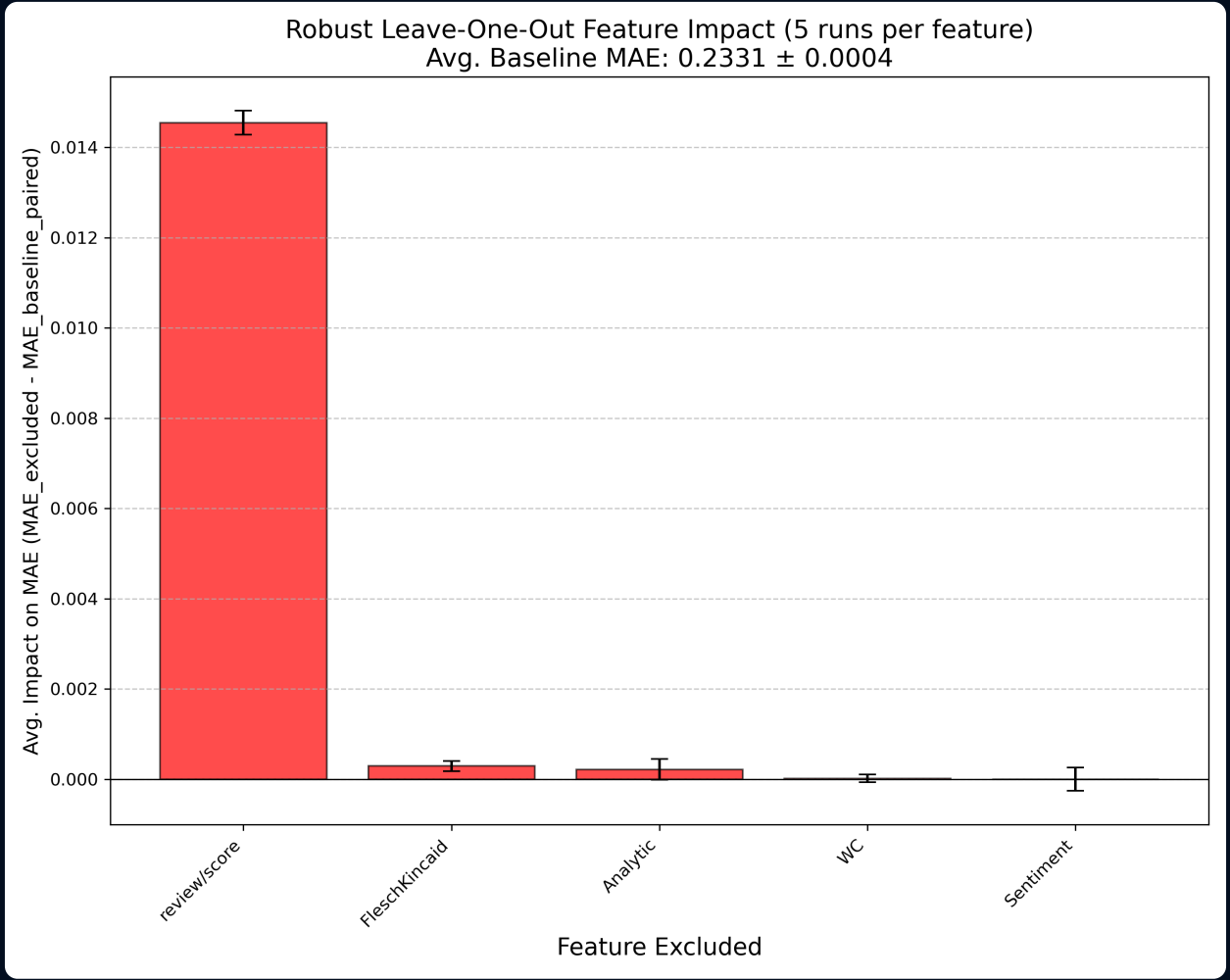
NEURAL NETWORK W/ EMPATH

RESULTS



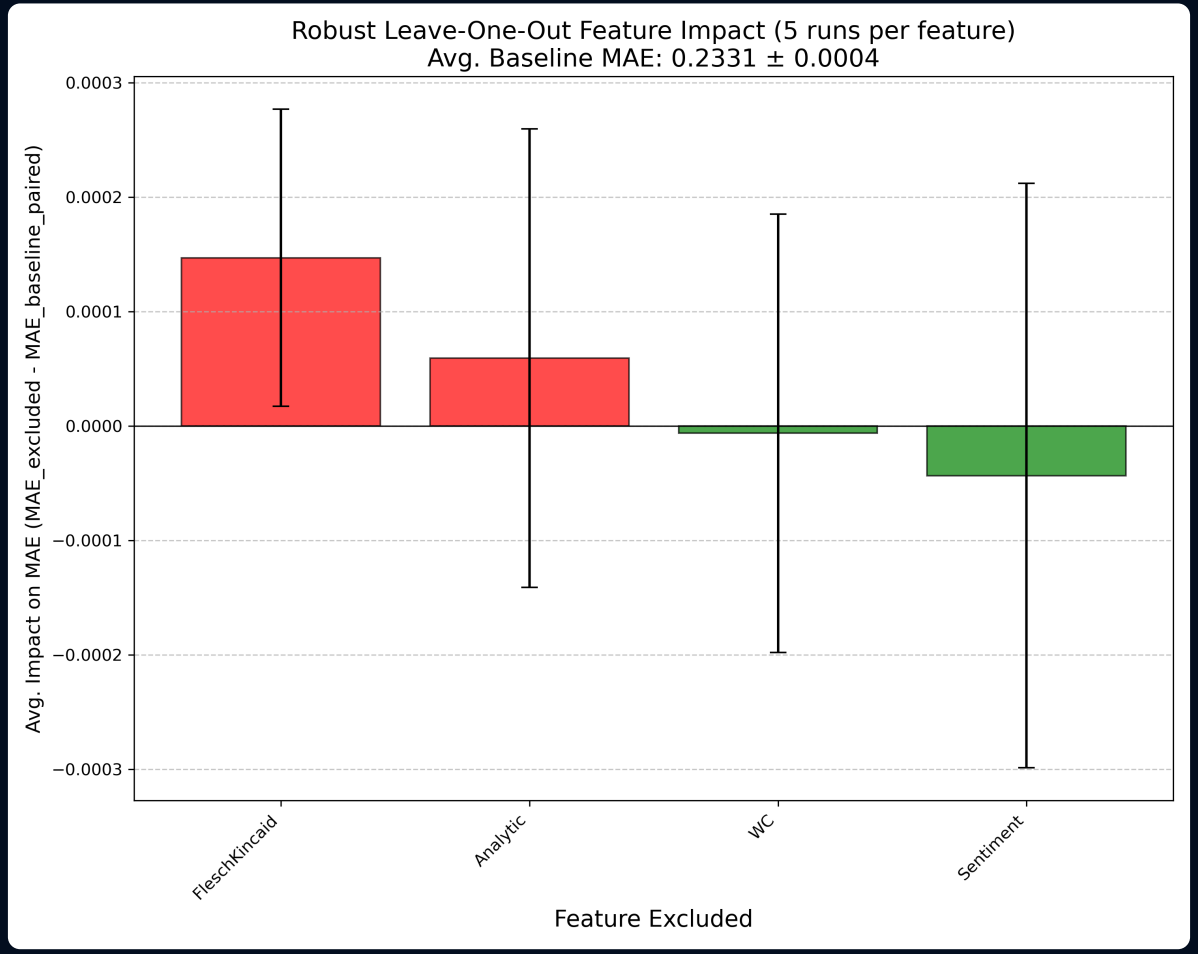
NEURAL NETWORK W/ EMPATH

FEATURE ANALYSIS



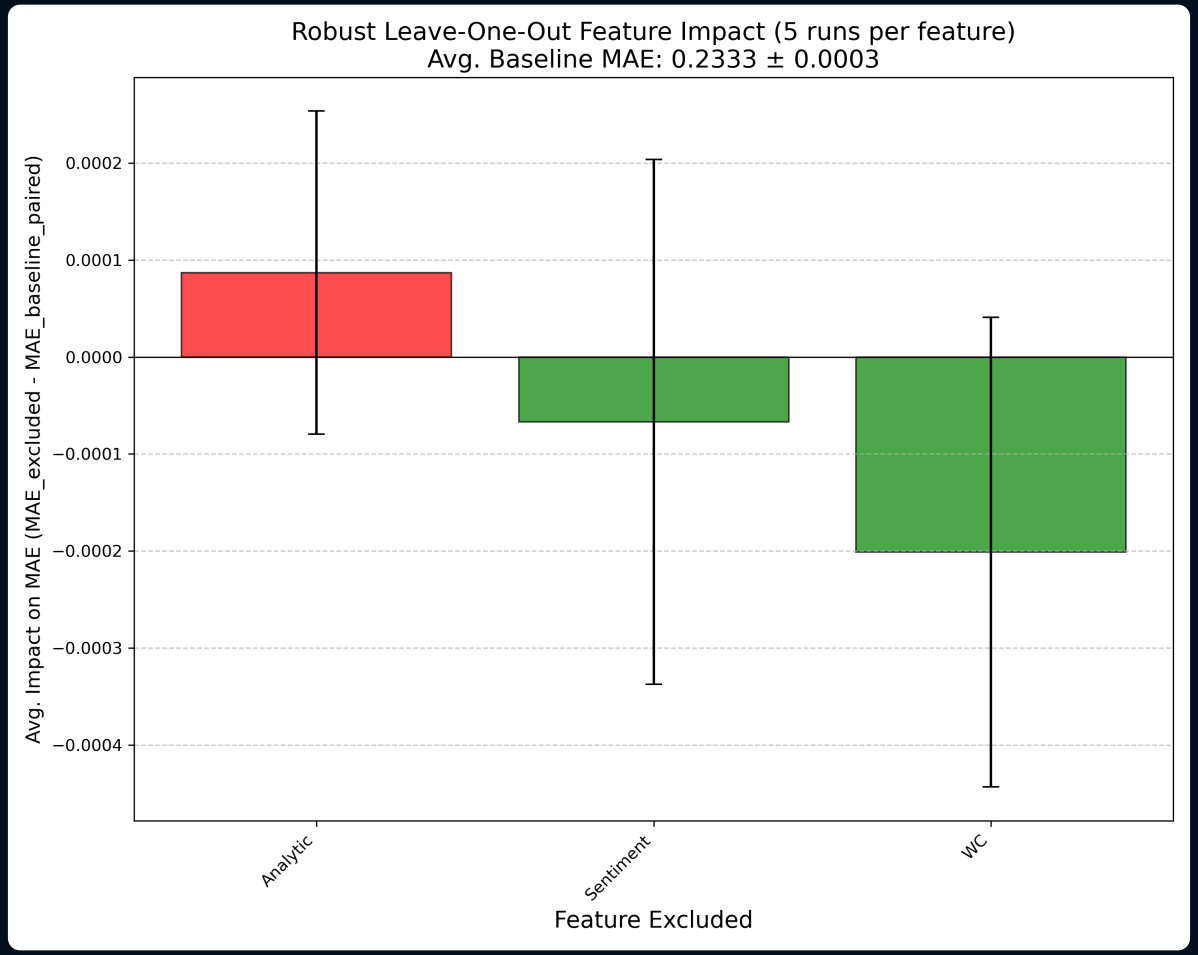
NEURAL NETWORK W/ EMPATH

FEATURE ANALYSIS



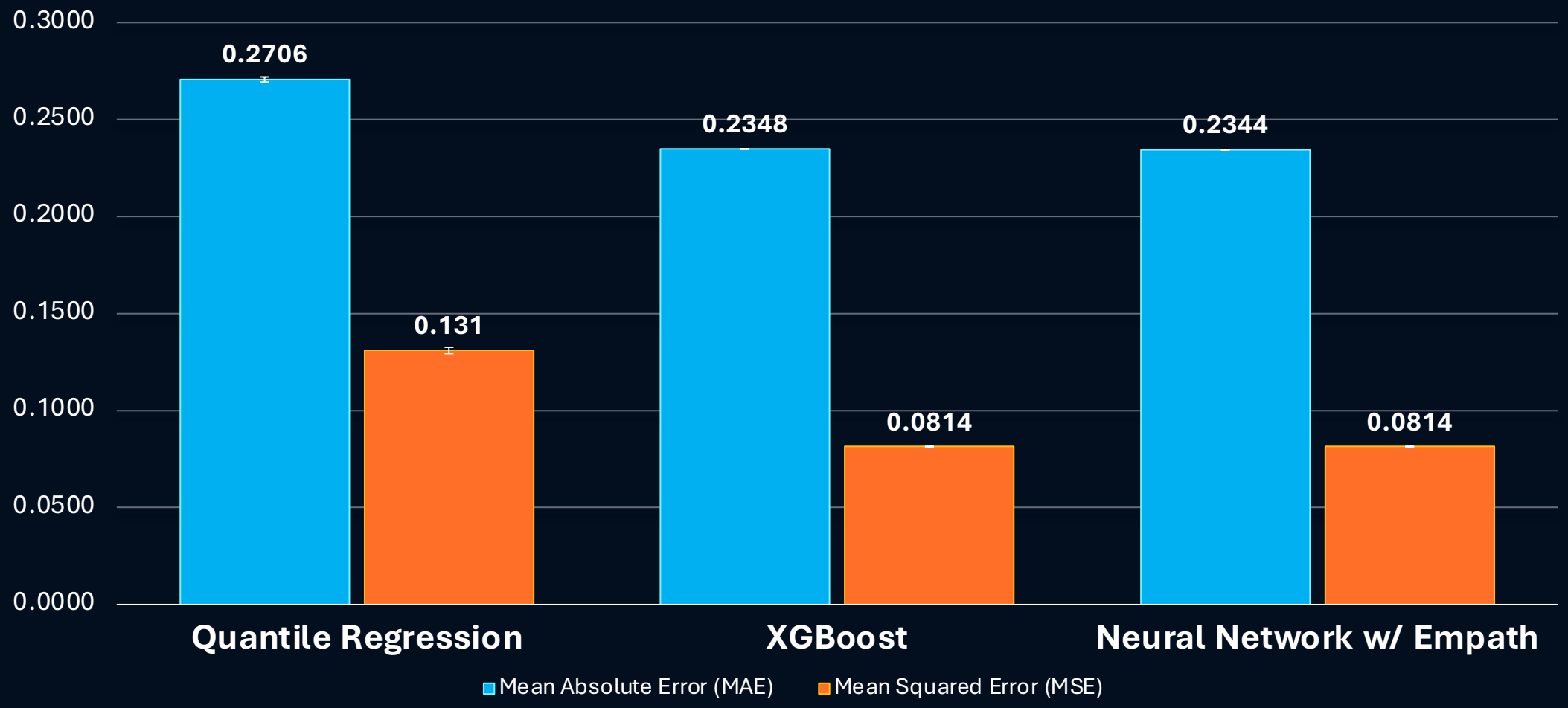
NEURAL NETWORK W/ EMPATH

FEATURE ANALYSIS



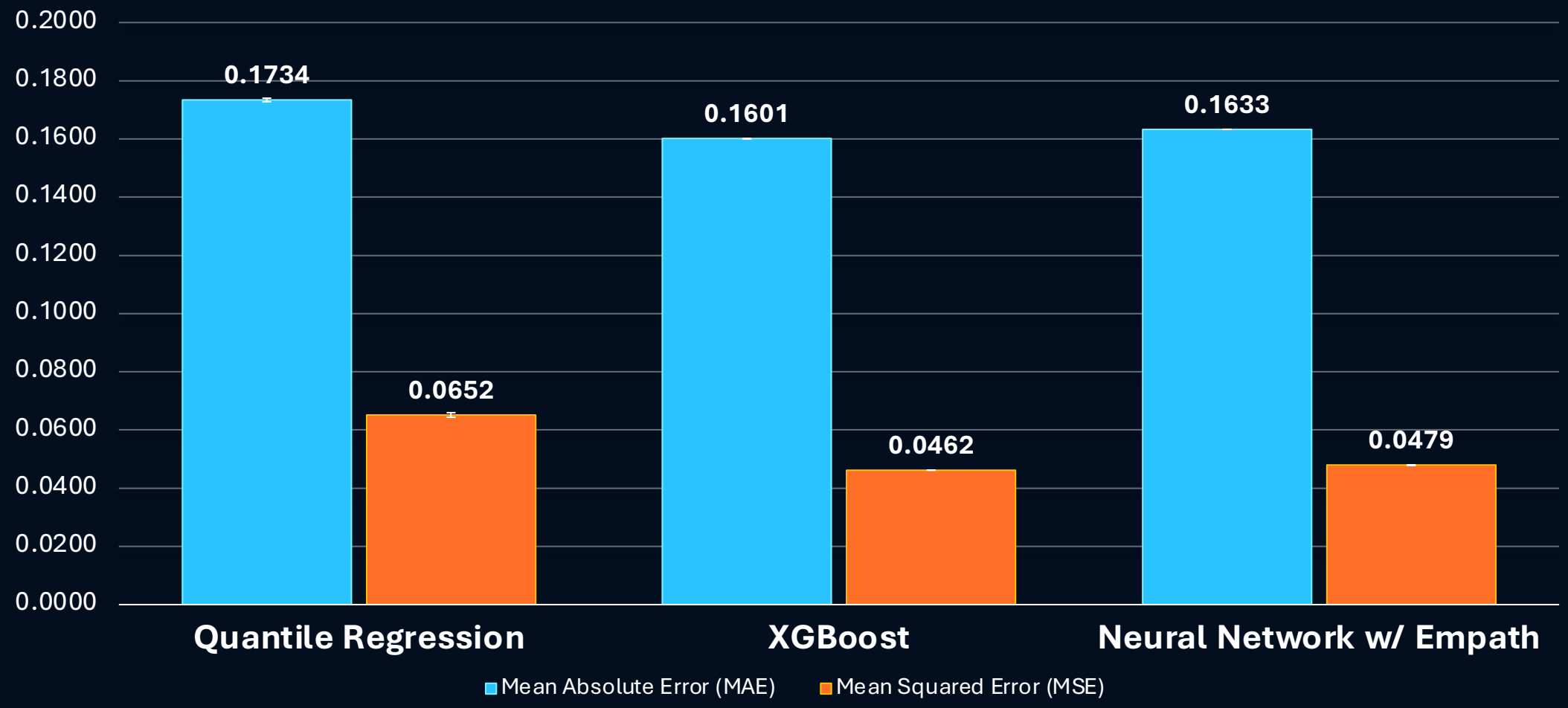
FULL DATASET

RESULTS



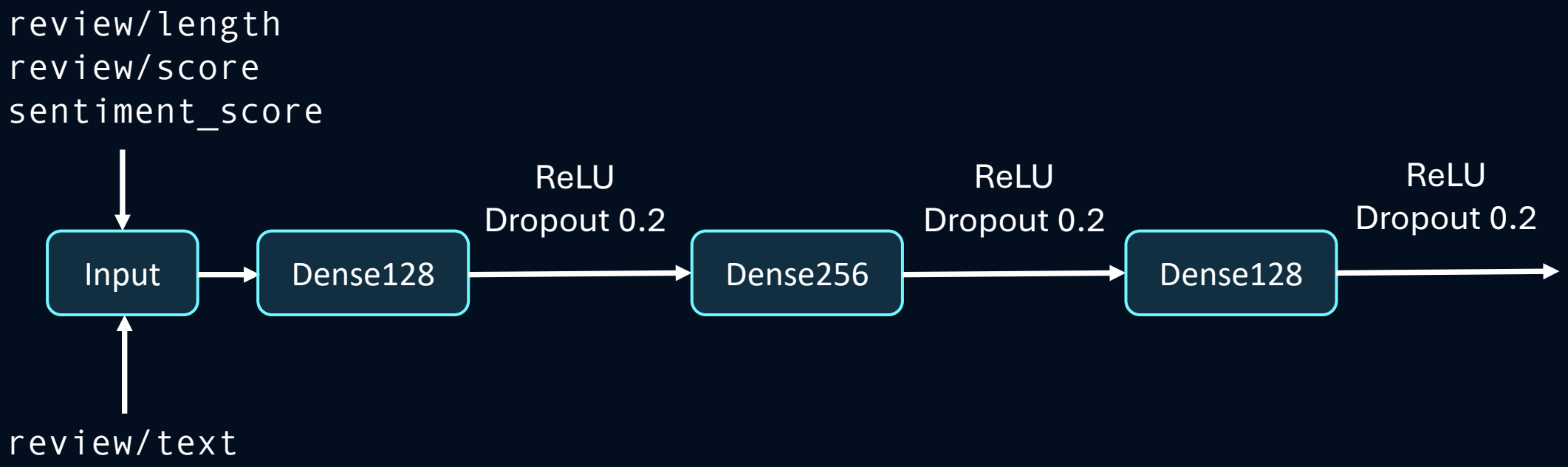
CERTAIN TWENTY

RESULTS



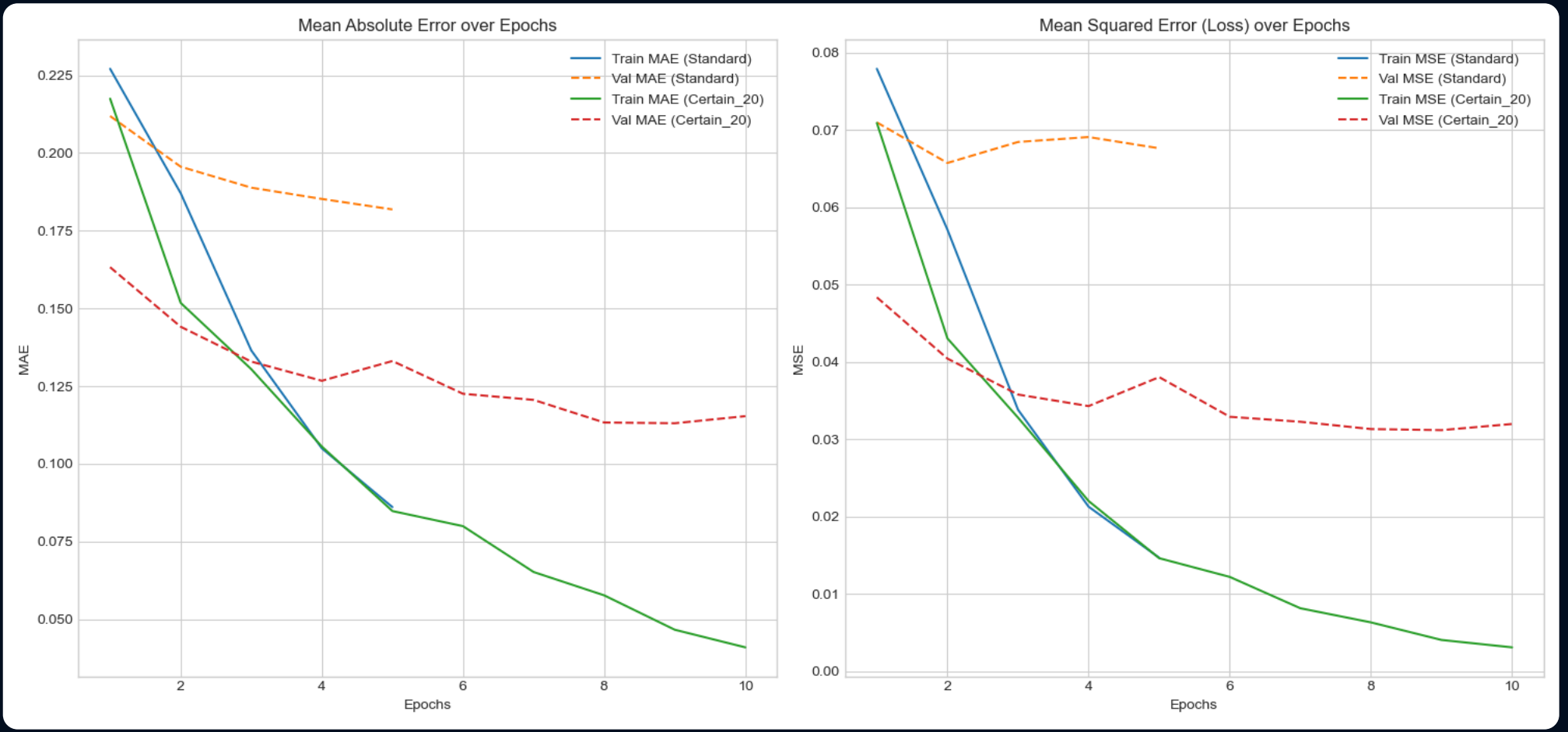
NEURAL NETWORK W/ TEXT

METHOD



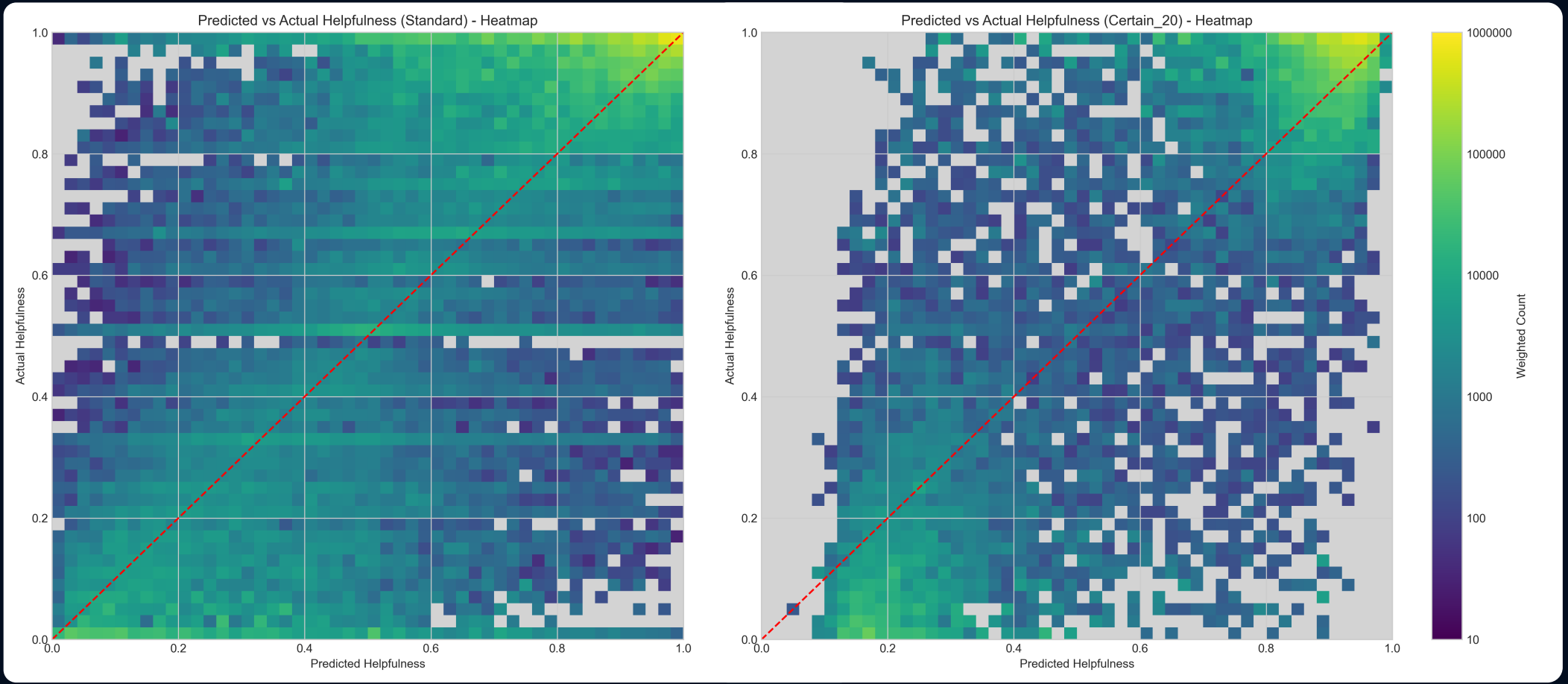
NEURAL NETWORK W/ TEXT

RESULTS



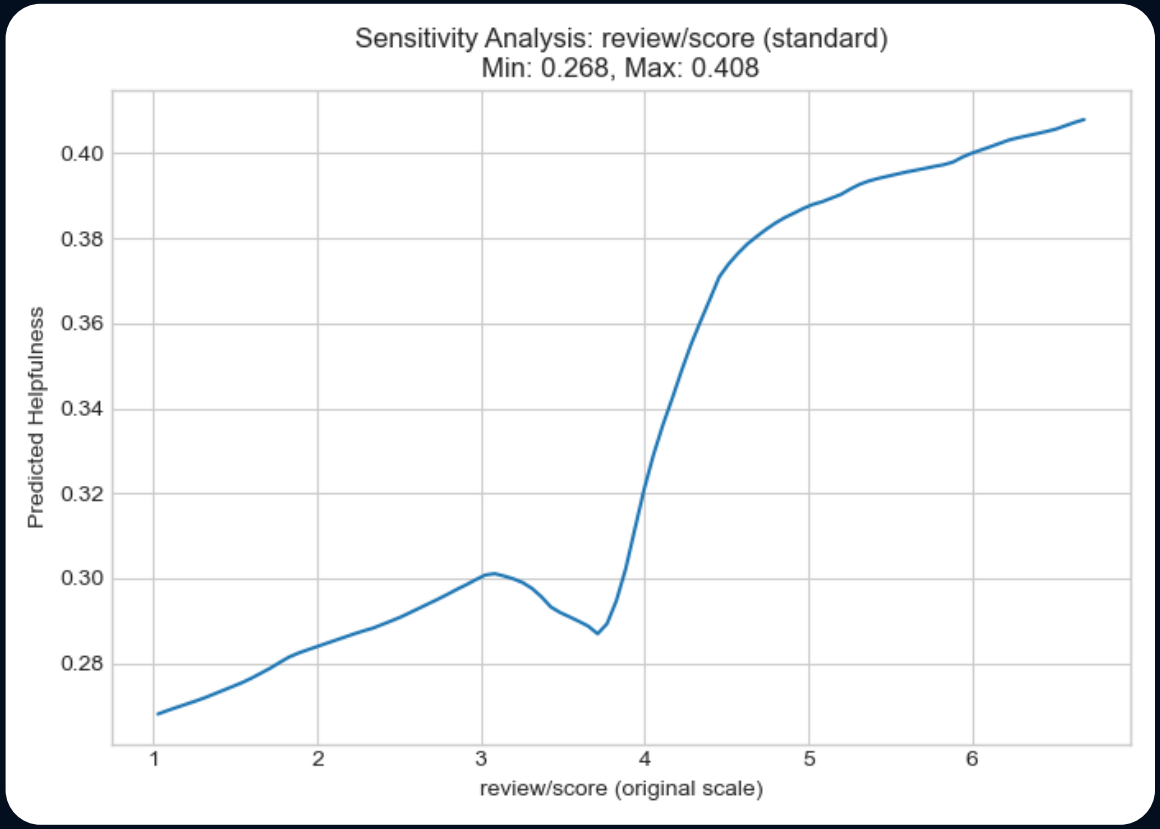
NEURAL NETWORK W/ TEXT

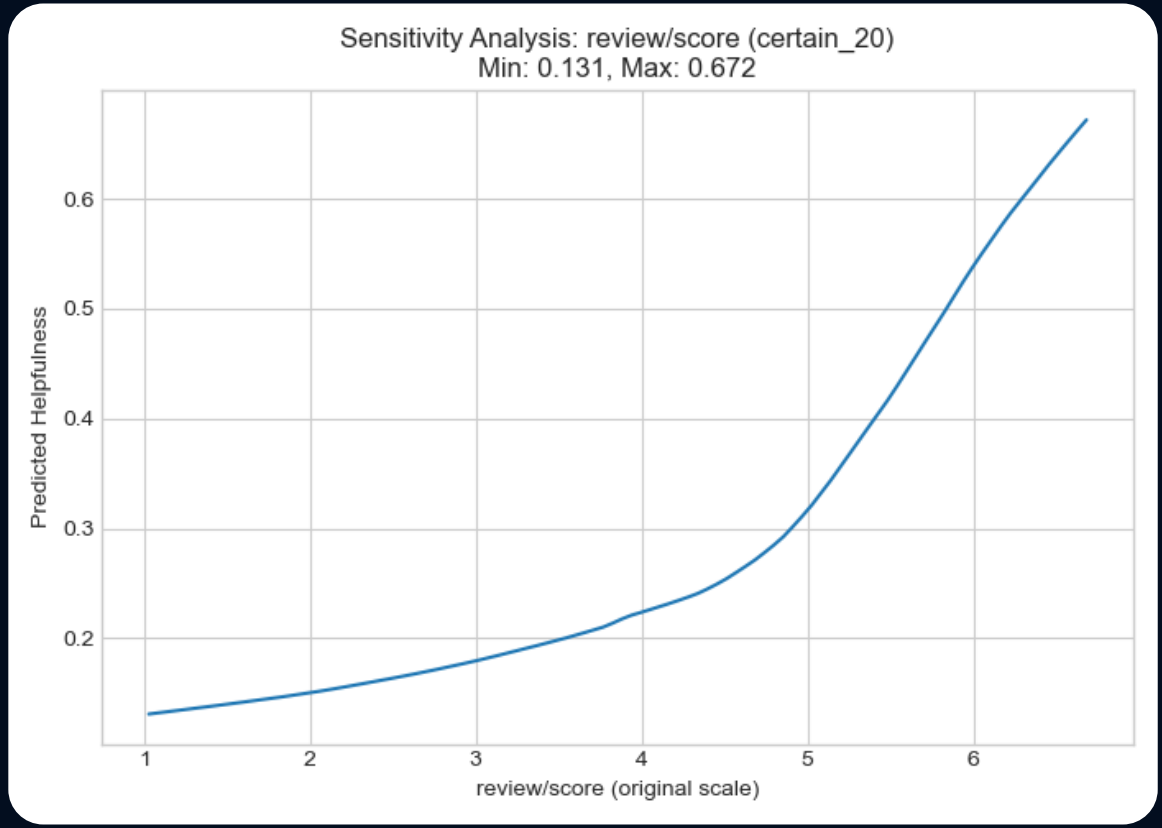
RESULTS



NN W/ TEXT: SENSITIVITY ANALYSIS

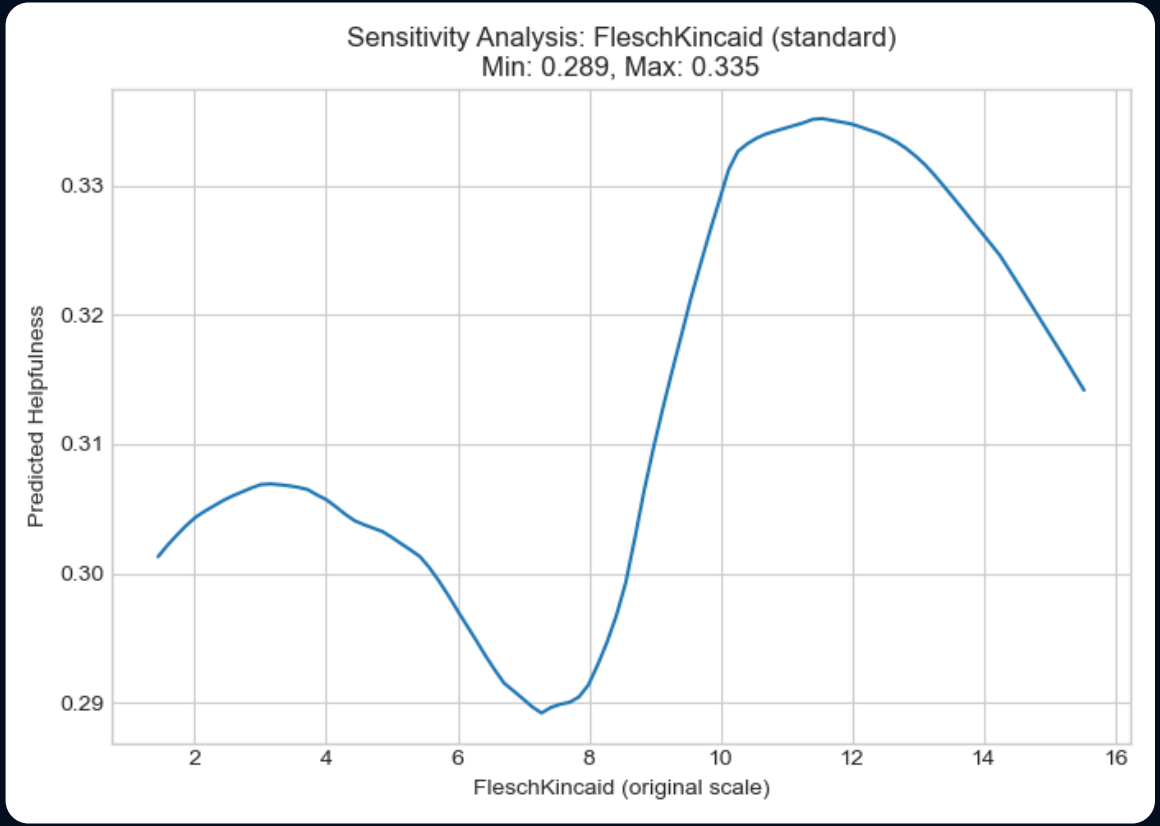
REVIEW/SCORE

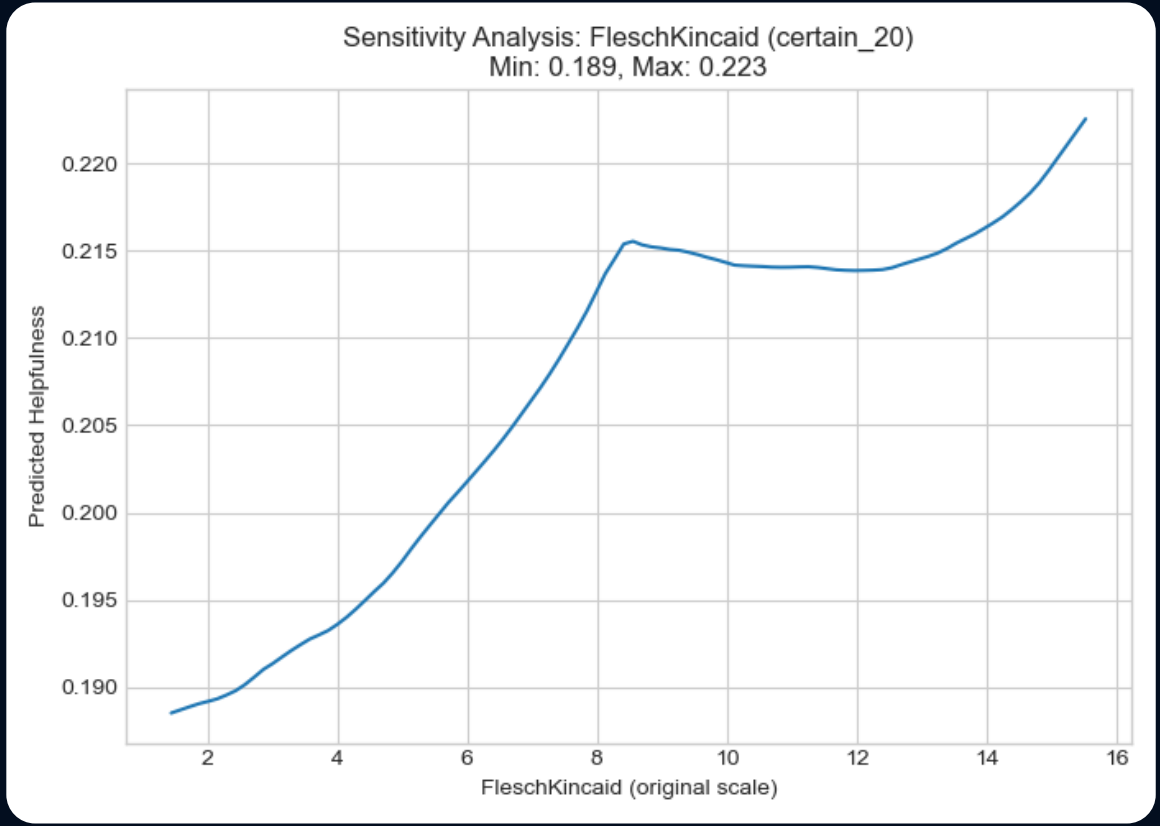




NN W/ TEXT: SENSITIVITY ANALYSIS

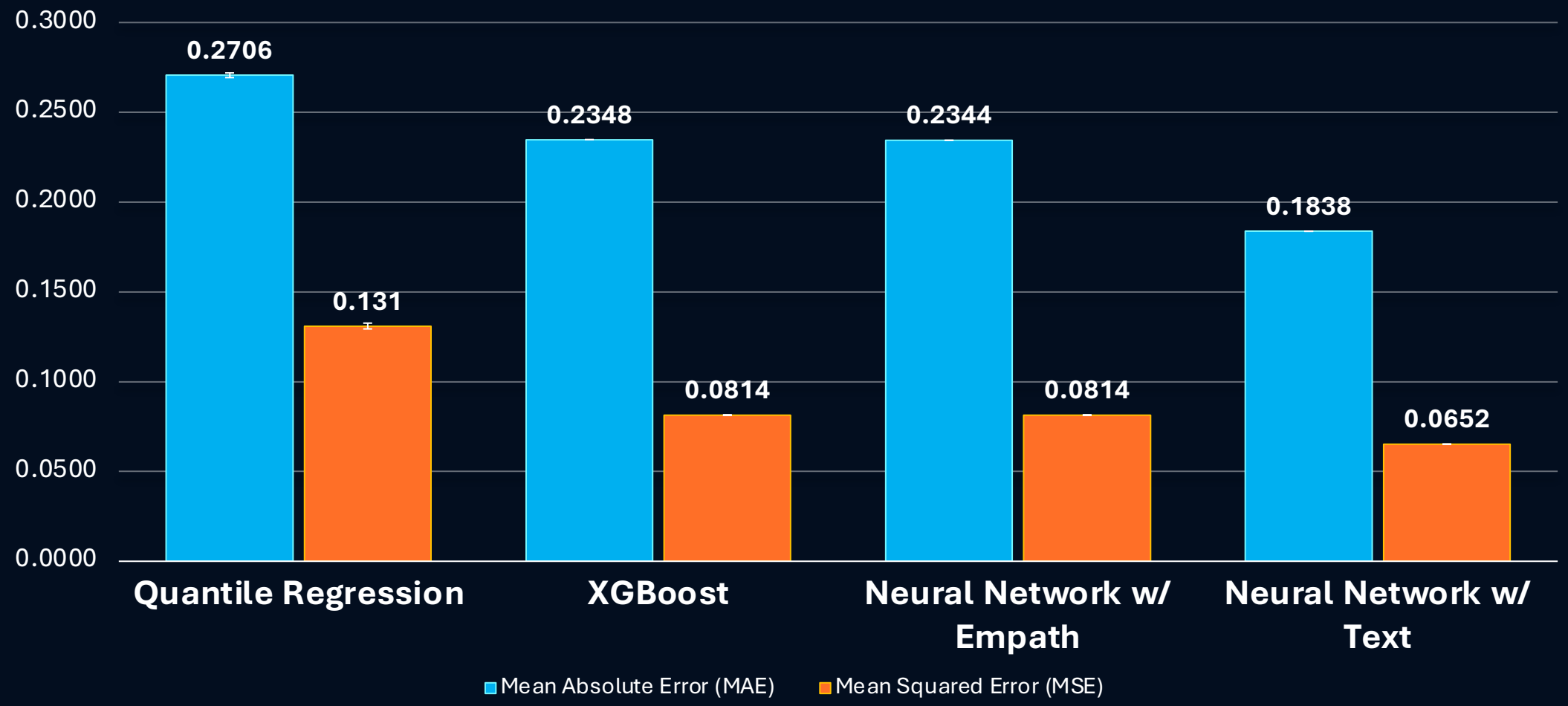
FLESCHKINCAID (READABILITY)





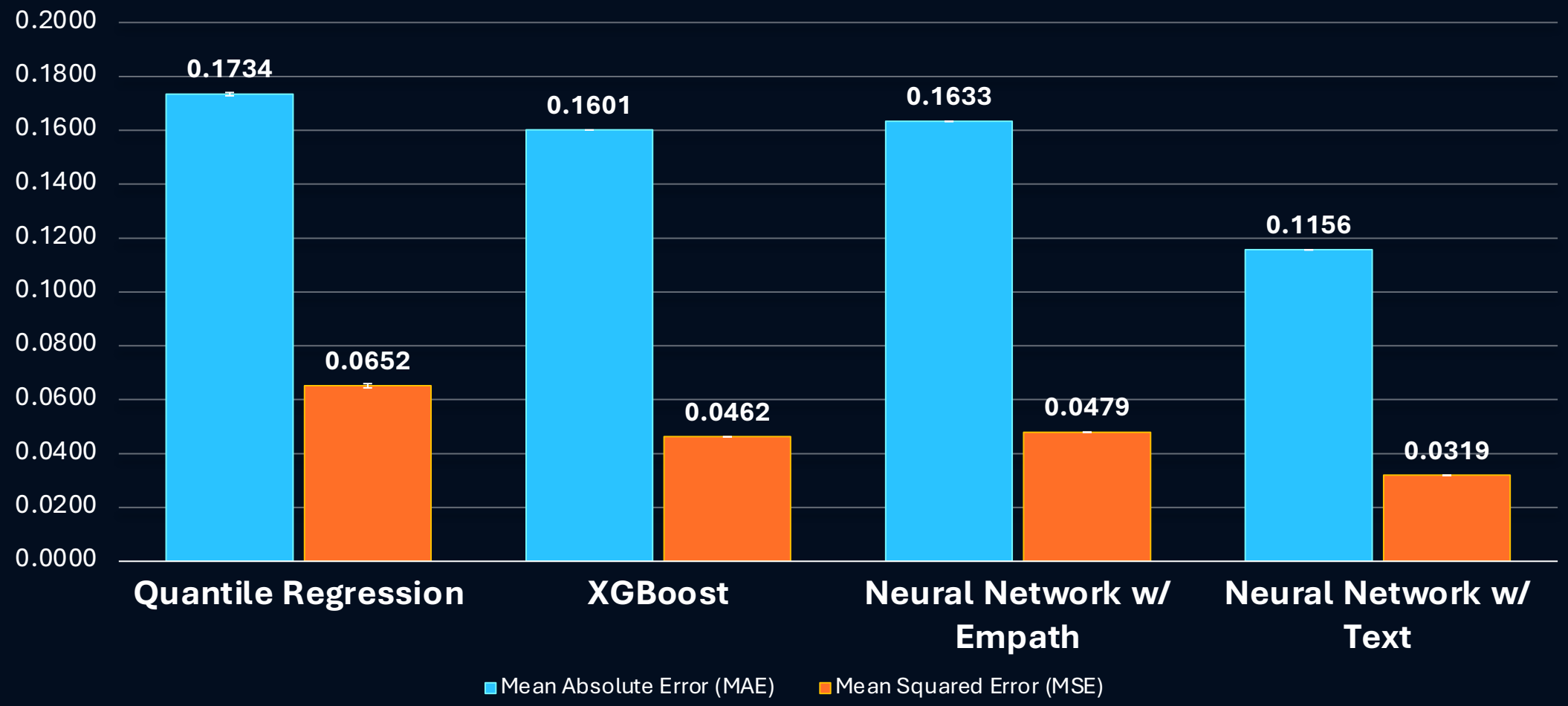
FULL DATASET

RESULTS



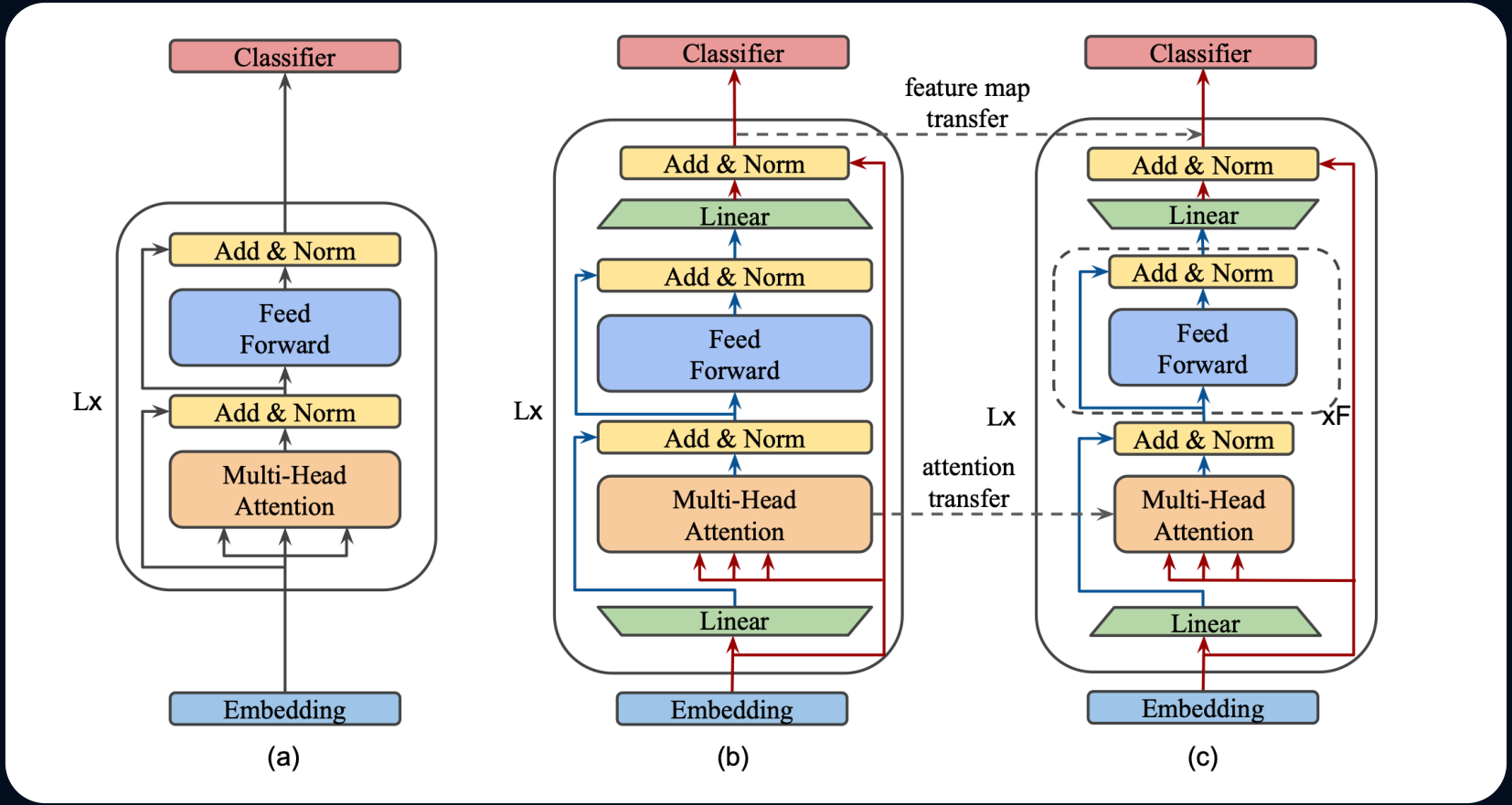
CERTAIN TWENTY

RESULTS



BERT

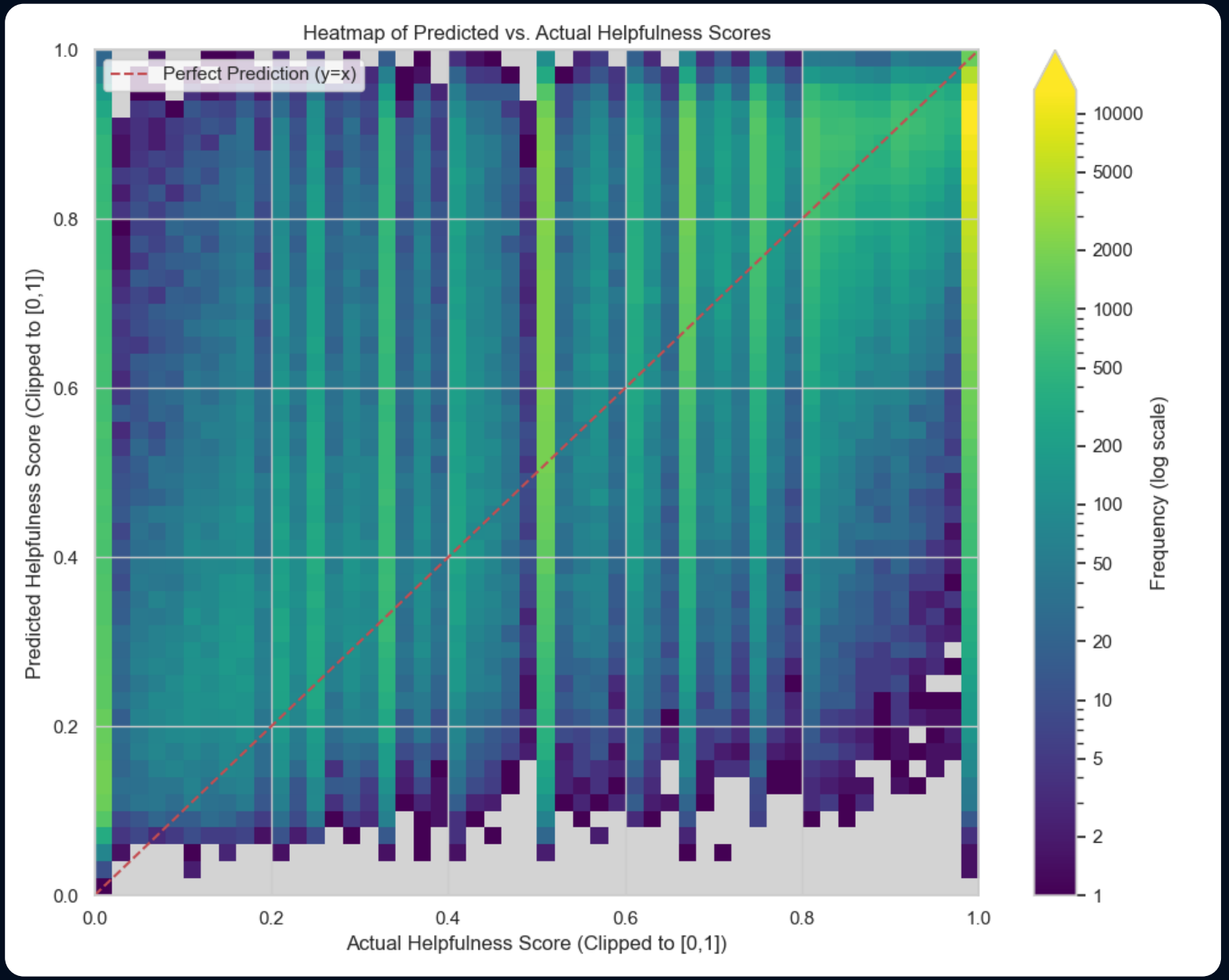
METHOD



https://arxiv.org/pdf/2004.02984v2 [9]

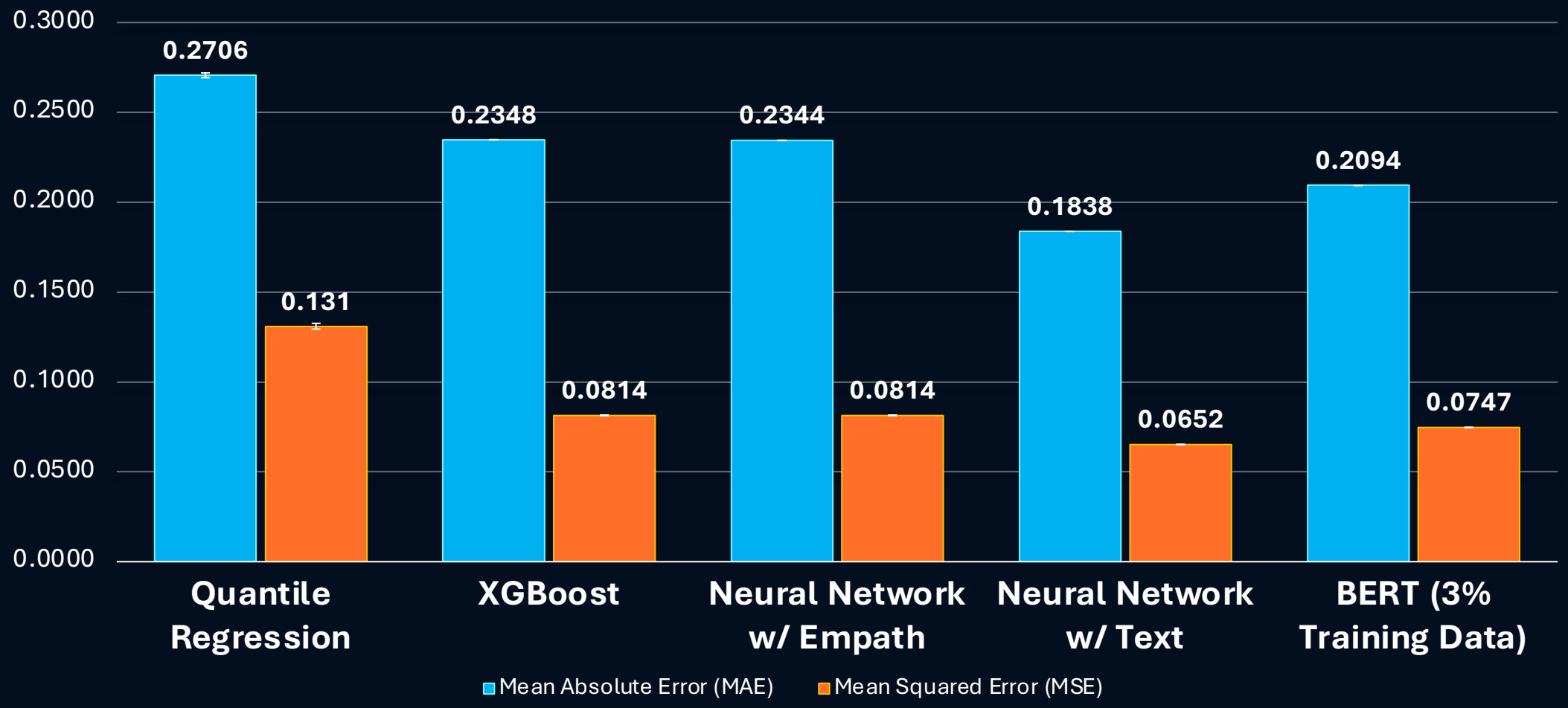
BERT

RESULTS



FULL DATASET

RESULTS

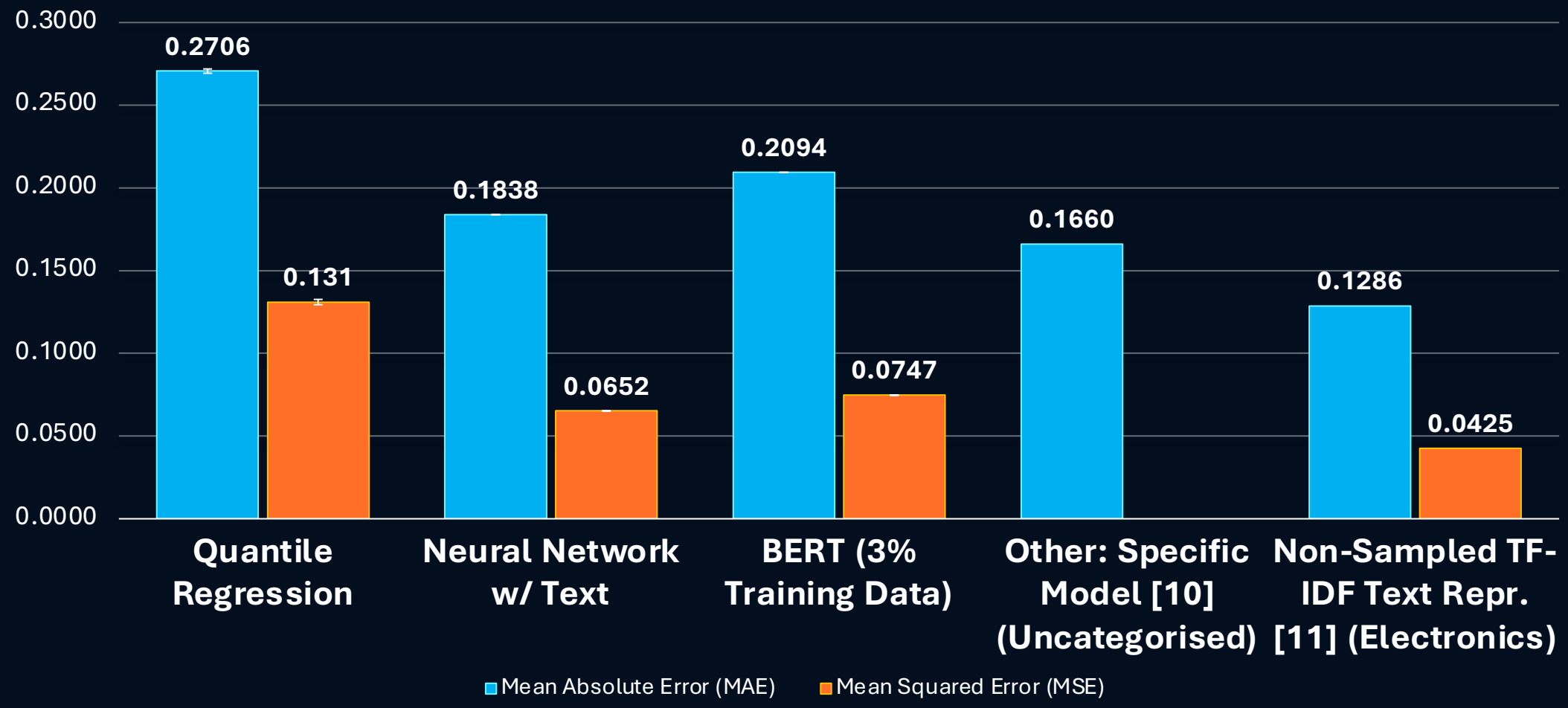


COMPARISON AND DISCUSSION

CONCLUSION

RESULTS COMPARISON

OTHER RESEARCH



DISCUSSION

AND LIMITATIONS

Our results have **large MAE / MSE** than previous research: however, important to note that this research was conducted on different Amazon Review Data Sets (not movies)

Only NN w/ Text, Quantile Regression, and BERT seem to avoid overfitting at edges

Using a **non-weighted** Quantile Regression for **non-weighted** MAE may yield different results

NN w/ Text seems like best model: however, **BERT** is in close second place, and was **only trained on 3% of the data**

FUTURE WORK

BERT was only trained on 3% of the dataset: increasing this will **potentially generate better results** than NN w/ Text model

Other smoothing methods than Laplace Smoothing could be implemented for the Quantile Regression model: these would likely have the largest effect for the Full Dataset (not Certain 20)

A method could be implemented to help minimise the effect of the **popularity-based rankings** on Amazon movie products

This investigation could be repeated with a dataset collected **closer to the present**, as the dataset utilised only runs until 2012

DECLARATION OF ORIGINALITY

WE, FERDINAND BRUNNE, OLIVER KING, BRIAN FUNK, AND SILVAN METZKER, DECLARE THAT THIS PROJECT IS ENTIRELY OUR OWN PIECE OF WORK.

REFERENCES

[1] Y. Park, “Predicting the Helpfulness of Online Customer Reviews across Different Product Types”, Sustainability, vol. 10(6), pp. 1735, May 2018.

[2] World Record Academy, “World's Largest Online Retailer and Marketplace”, https://www.worldrecordacademy.org/2024/11/worlds-largest-online-retailer-and-marketplace-world-record-set-by-amazon-424452, Nov. 2024. [Accessed 11 May 2025]

[3] D. Parris, “The Rise, Fall, and (Slight) Rise of DVDs. A Statistical Analysis”, https://www.statsignificant.com/p/the-rise-fall-and-slight-rise-of, Dec. 2023. [Accessed 11 May 2025]

[4] J. Leskovec, “Web data: Amazon Movie Reviews”, https://snap.stanford.edu/data/web-Movies.html, May. 2013. [Accessed 11 May 2025]

[5] S. Kim et al, “Automatically assessing review helpfulness”, EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 423-430, July 2006.

[6] A. Ghose and P. G. Ipeirotis, “Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics”, IEEE Transactions on Knowledge and Data Engineering, vol. 23(10), pp. 1498-1512, Oct. 2011.

REFERENCES

[7] M. Bilal and A. A. Almazroi, “Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews”, Electron Commer Res, vol.23, pp. 2737-2757, Dec. 2023.

[8] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, https://arxiv.org/abs/1603.02754, Mar. 2016. [Accessed 11/05/2025]

[9] Z. Sun et al, “MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices”, https://arxiv.org/pdf/2004.02984v2, Apr. 2020. [Accessed 11/05/2025]

[10] K. M. Shannon, “Predicting Amazon Review Helpfulness Ratio”, https://www.researchgate.net/publication/317578901_Predicting_Amazon_Review_Helpfulness_Ratio, June 2017. [Accessed 11/05/2025]

[11] F. Hjalmarsson, “Predicting the Helpfulness of Online Product Reviews”, https://www.diva-portal.org/smash/get/diva2:1595730/FULLTEXT02.pdf, Sept. 2021. [Accessed 11/05/2025]

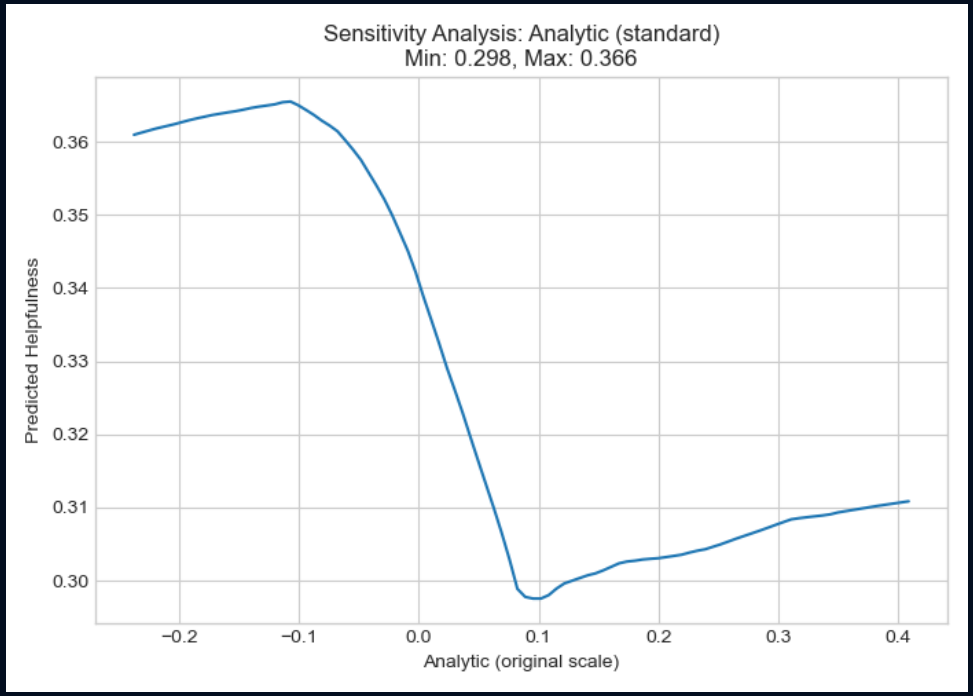
THANK YOU FOR LISTENING

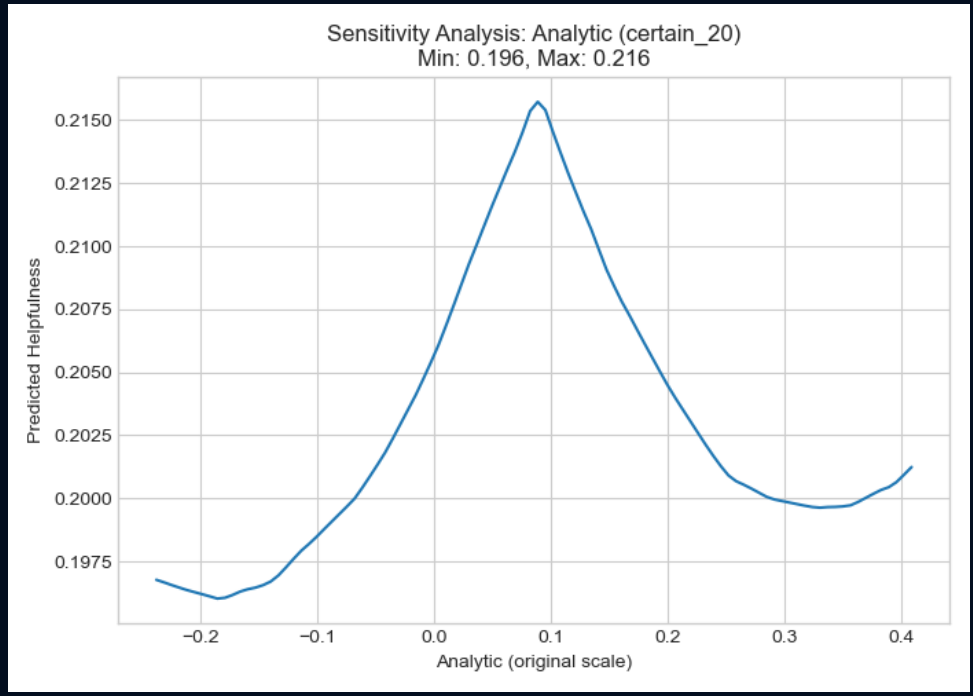
ANY QUESTIONS ?

EXTRA SLIDES

ANALYSIS

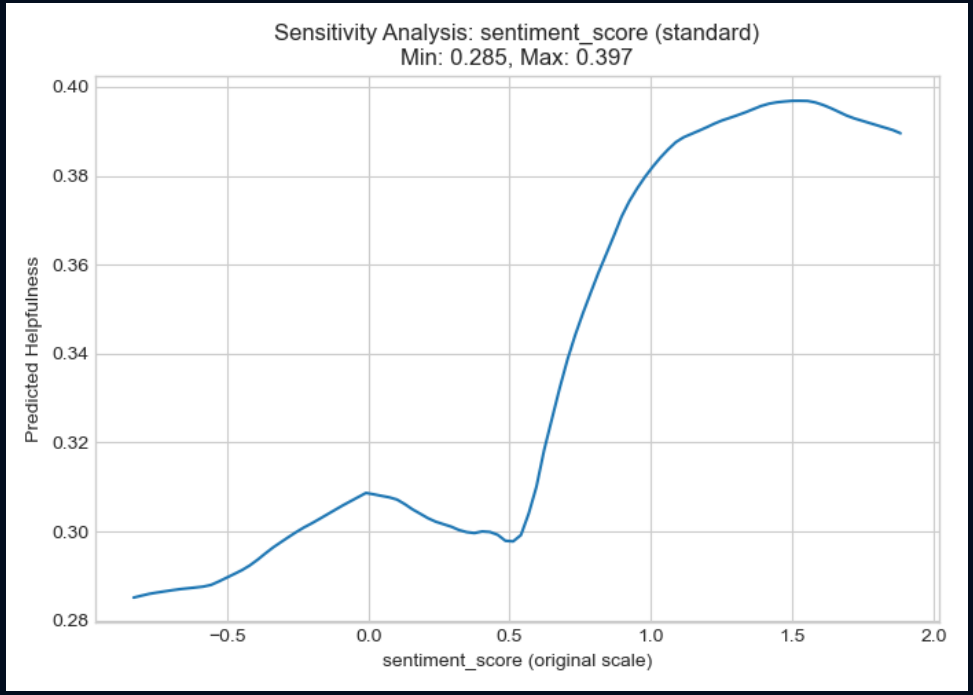
ANALYTIC

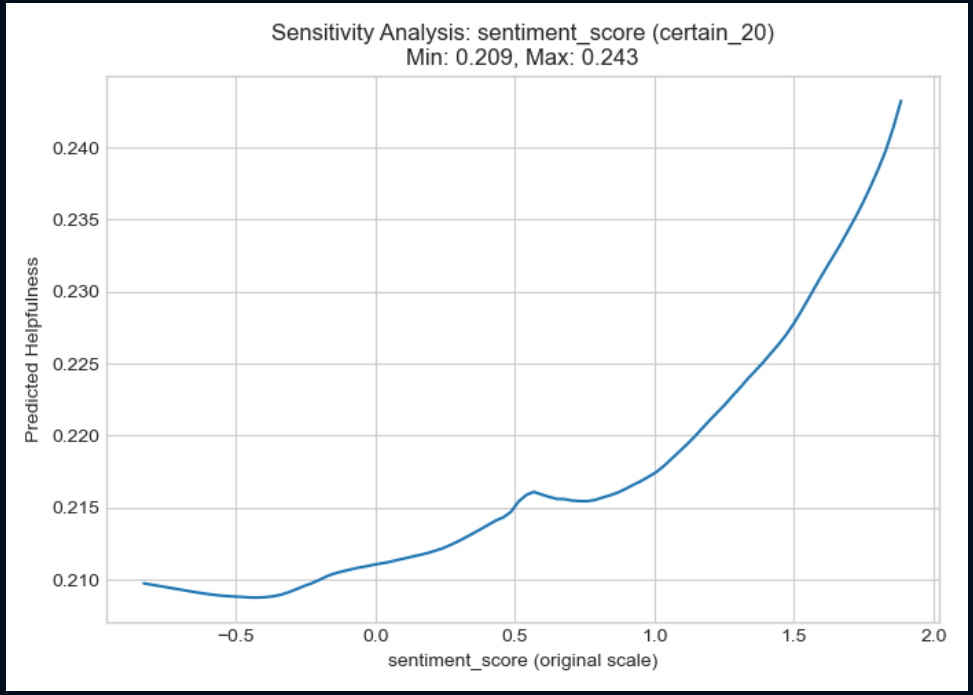




ANALYSIS

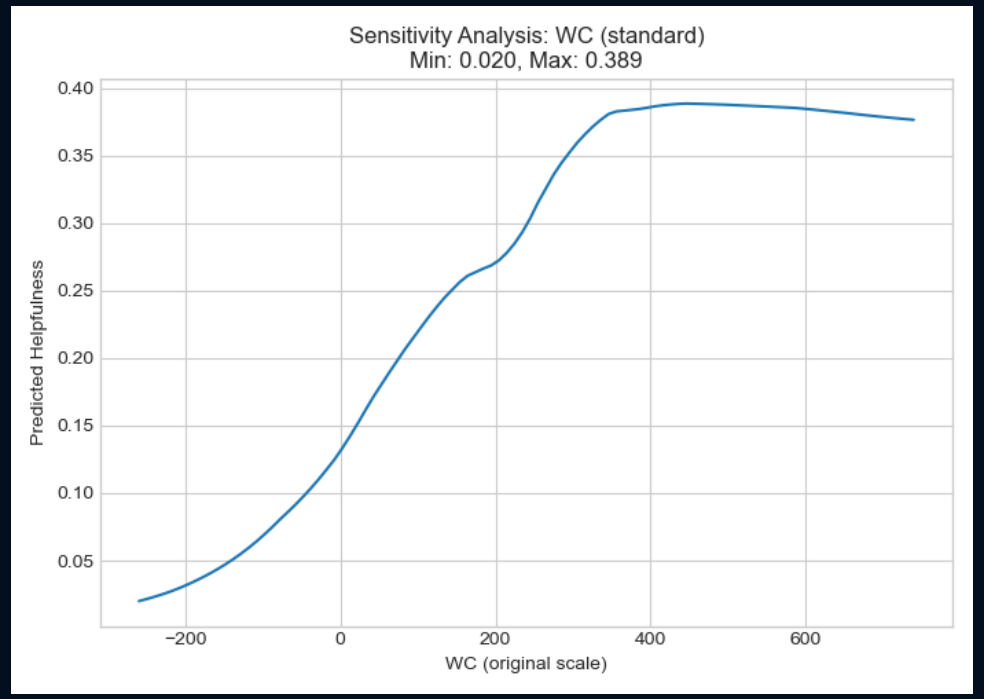
SENTIMENT SCORE

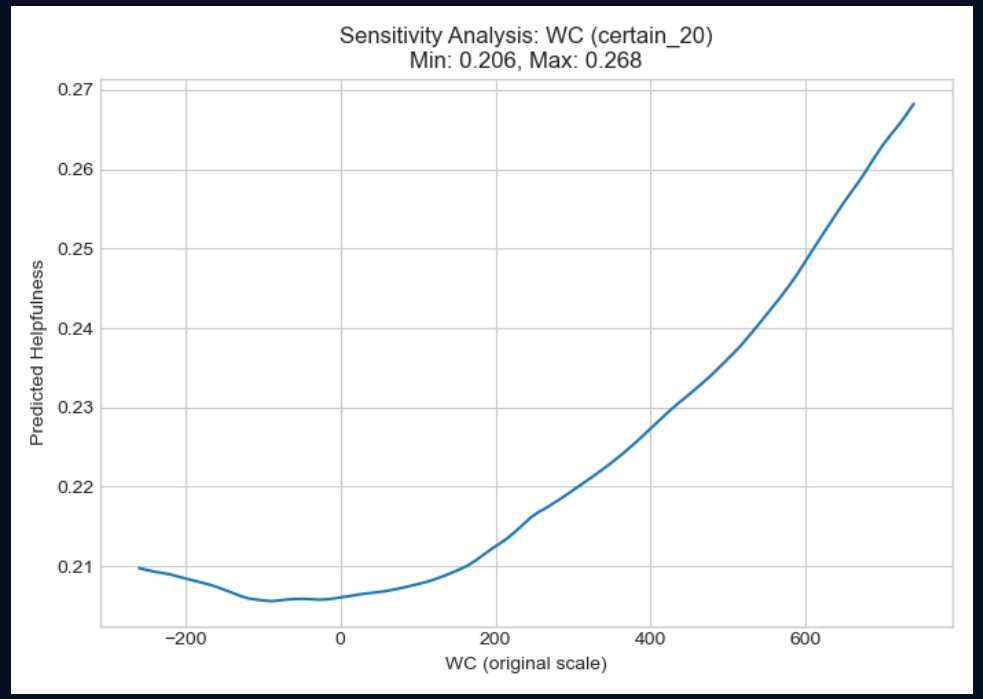




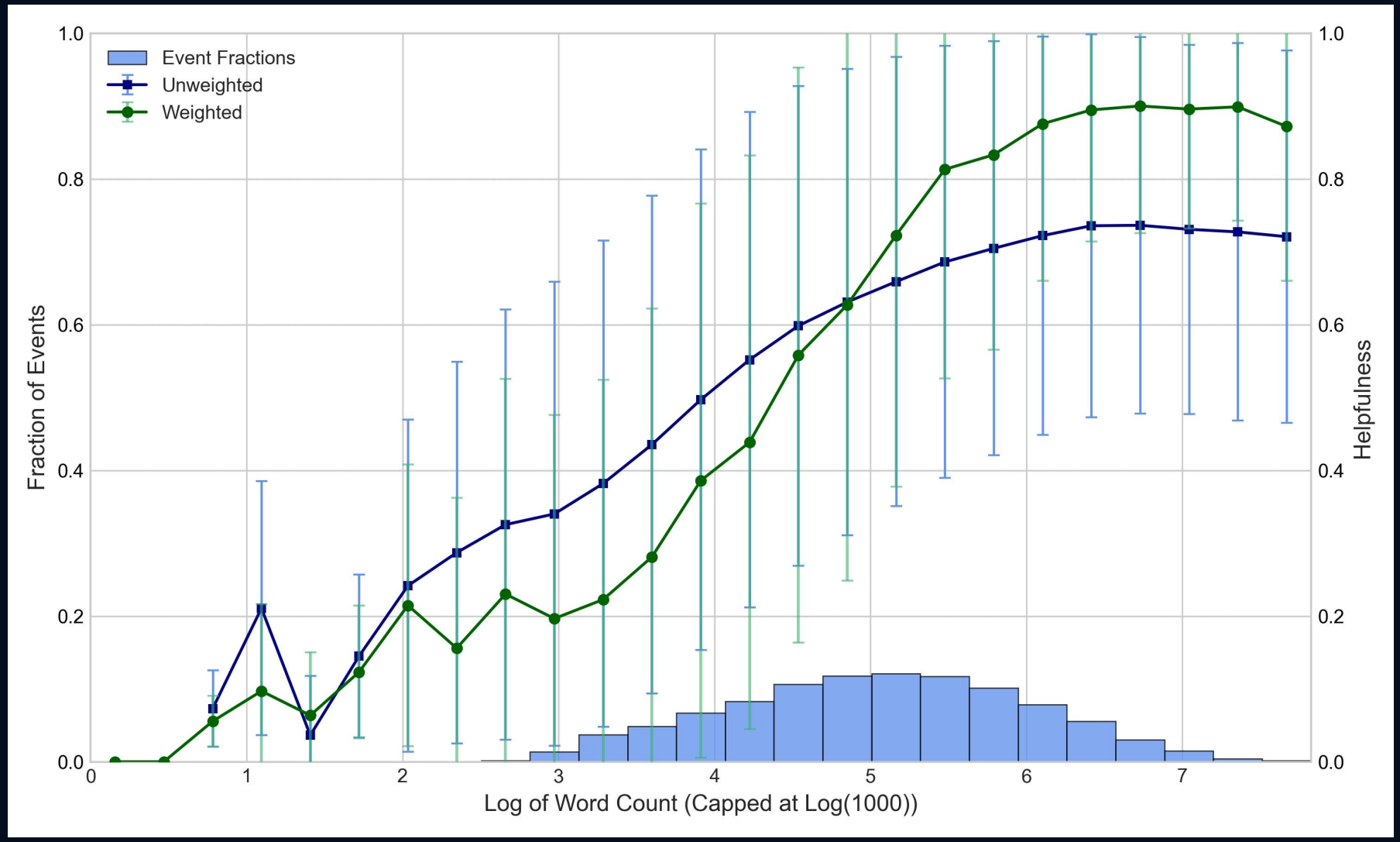
ANALYSIS

WORD COUNT

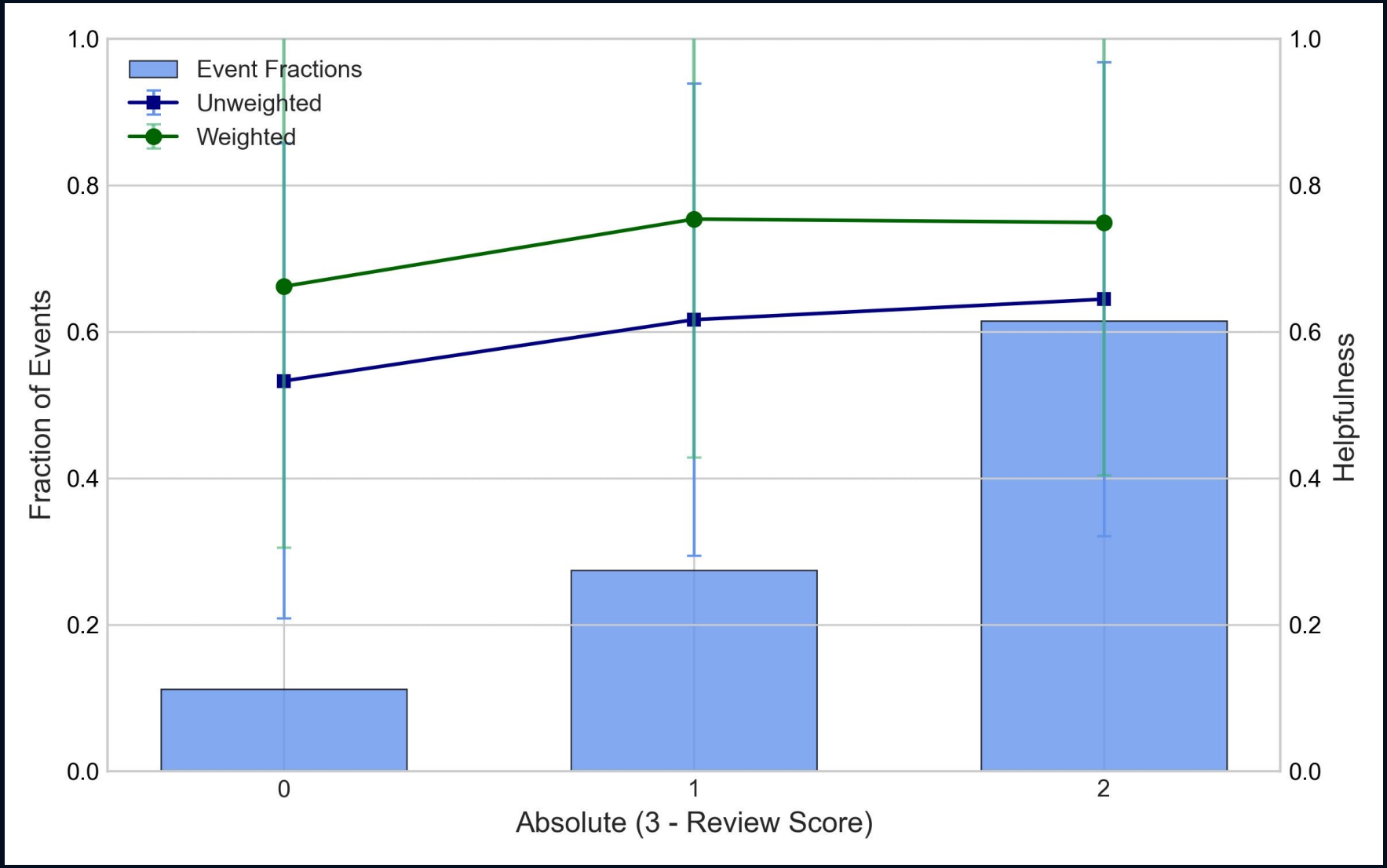




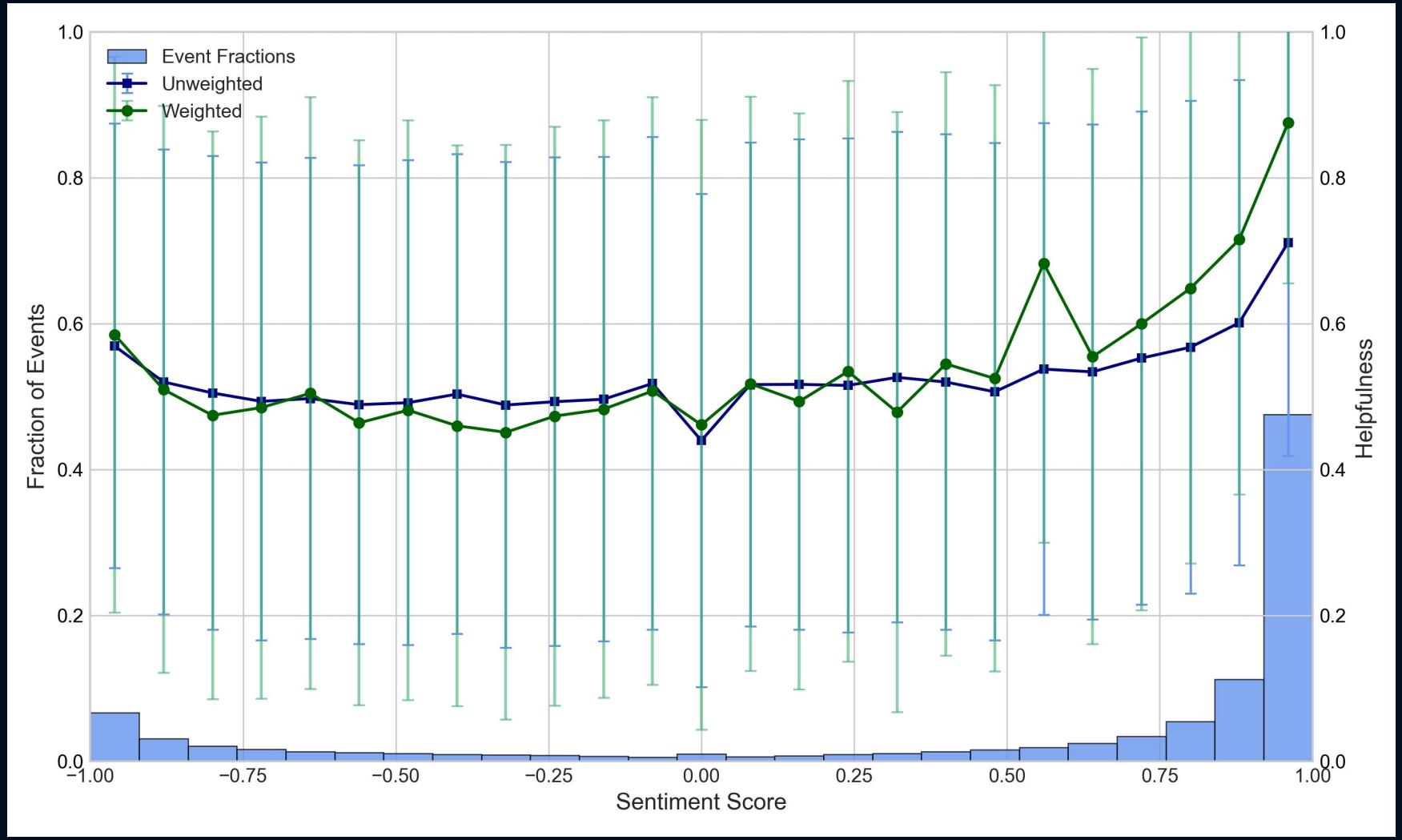
EDA: LOG(WORD COUNT)



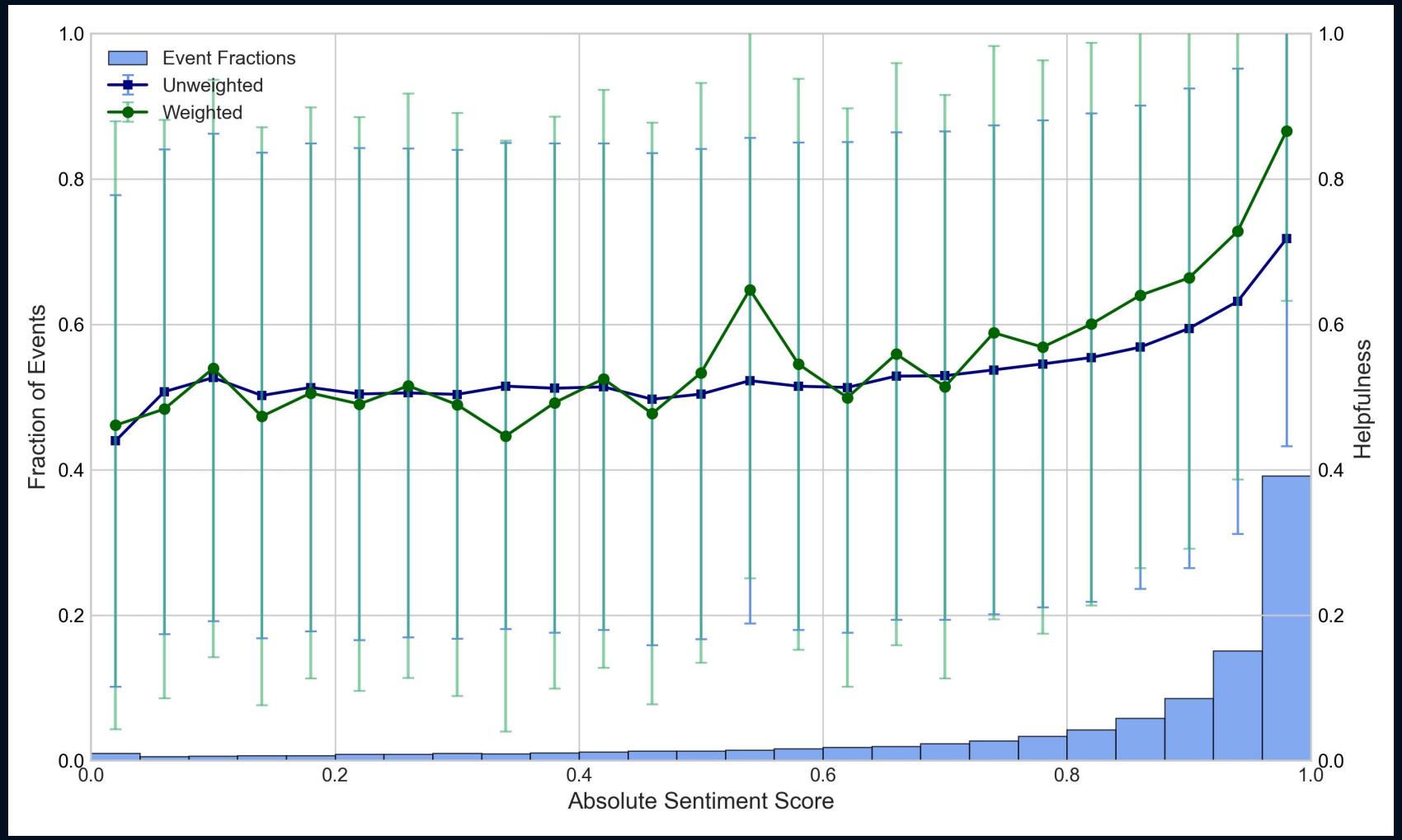
EDA: ABS(THREE – RATING)



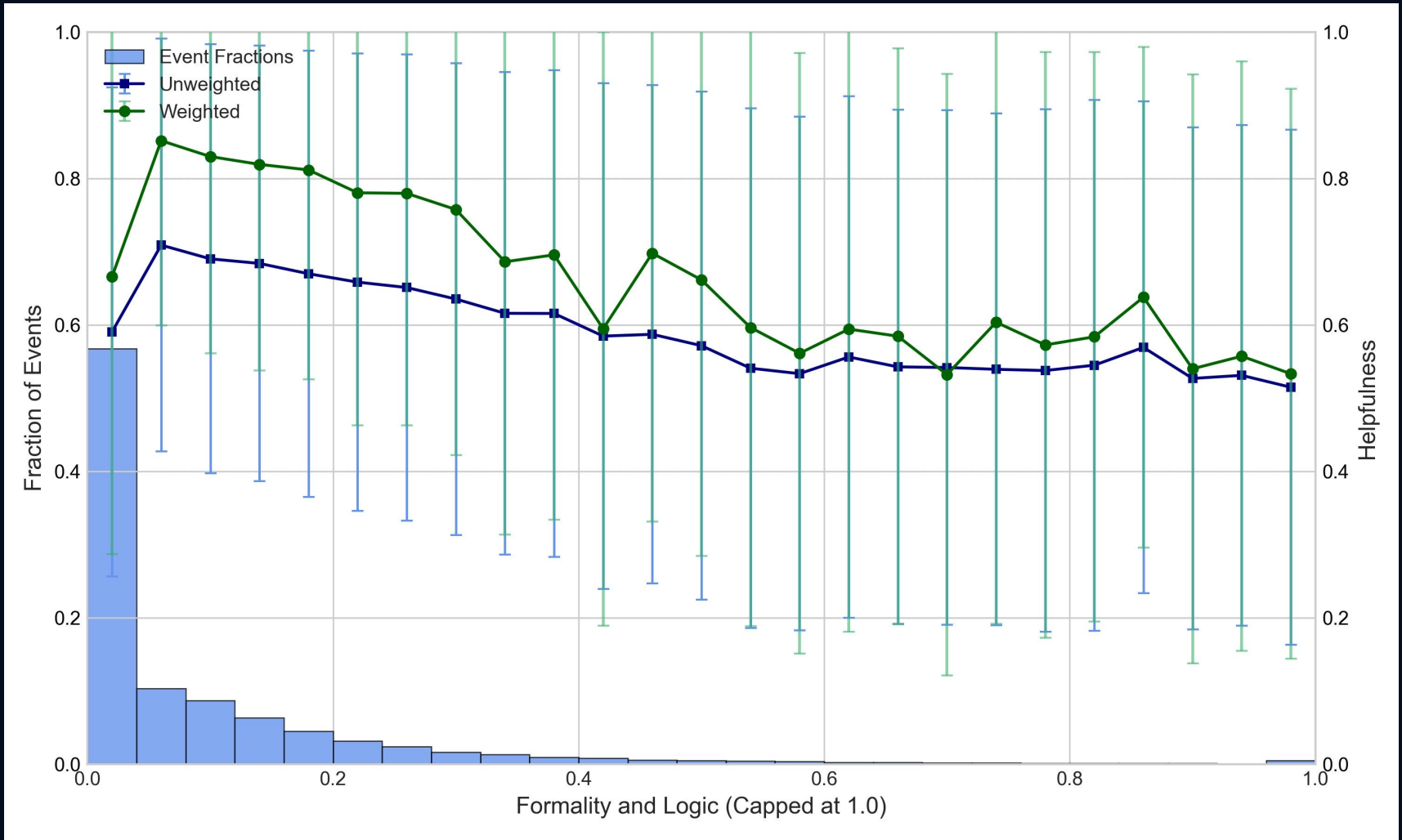
EDA: SENTIMENT



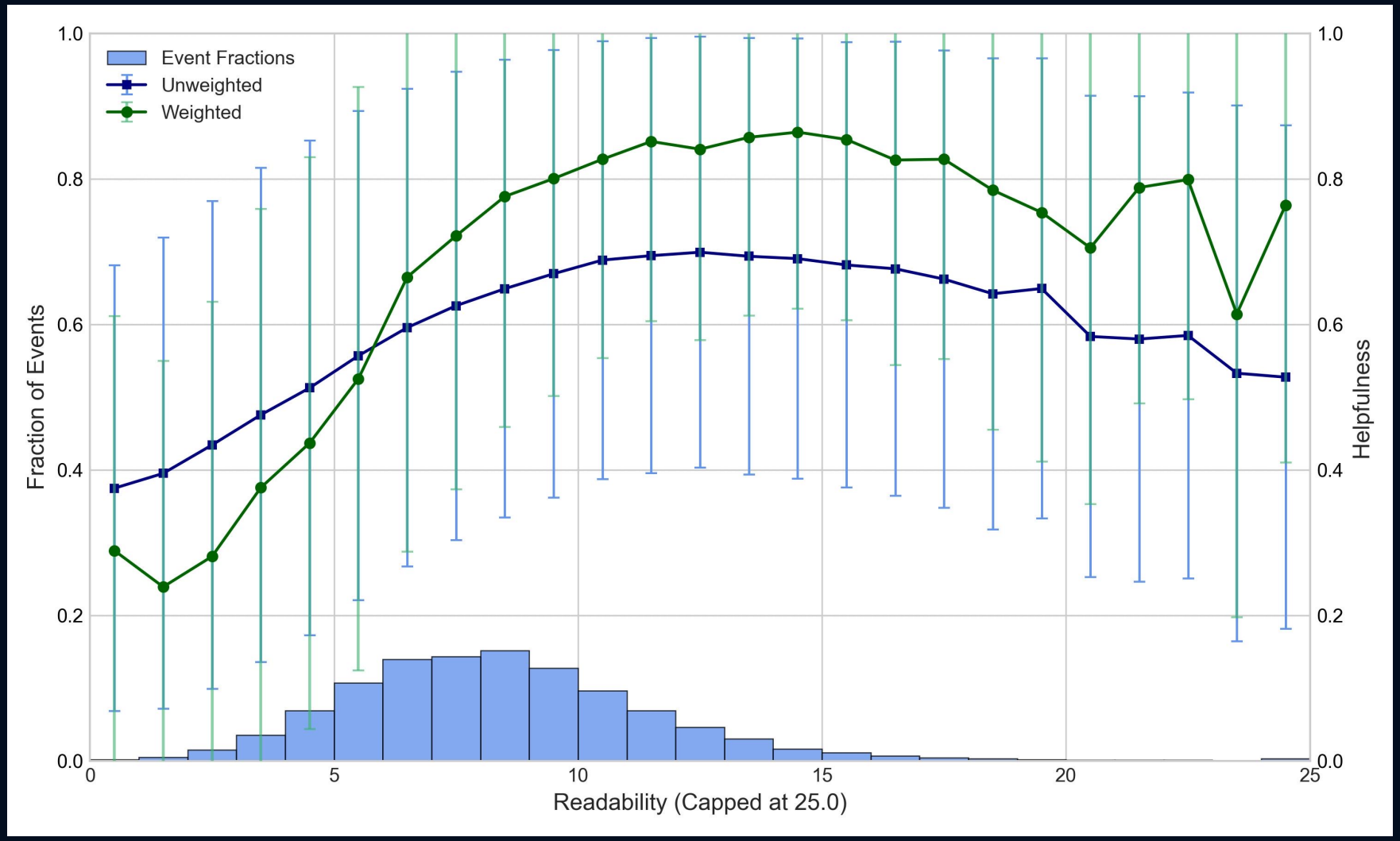
EDA: ABS(SENTIMENT)



EDA: ANALYTIC

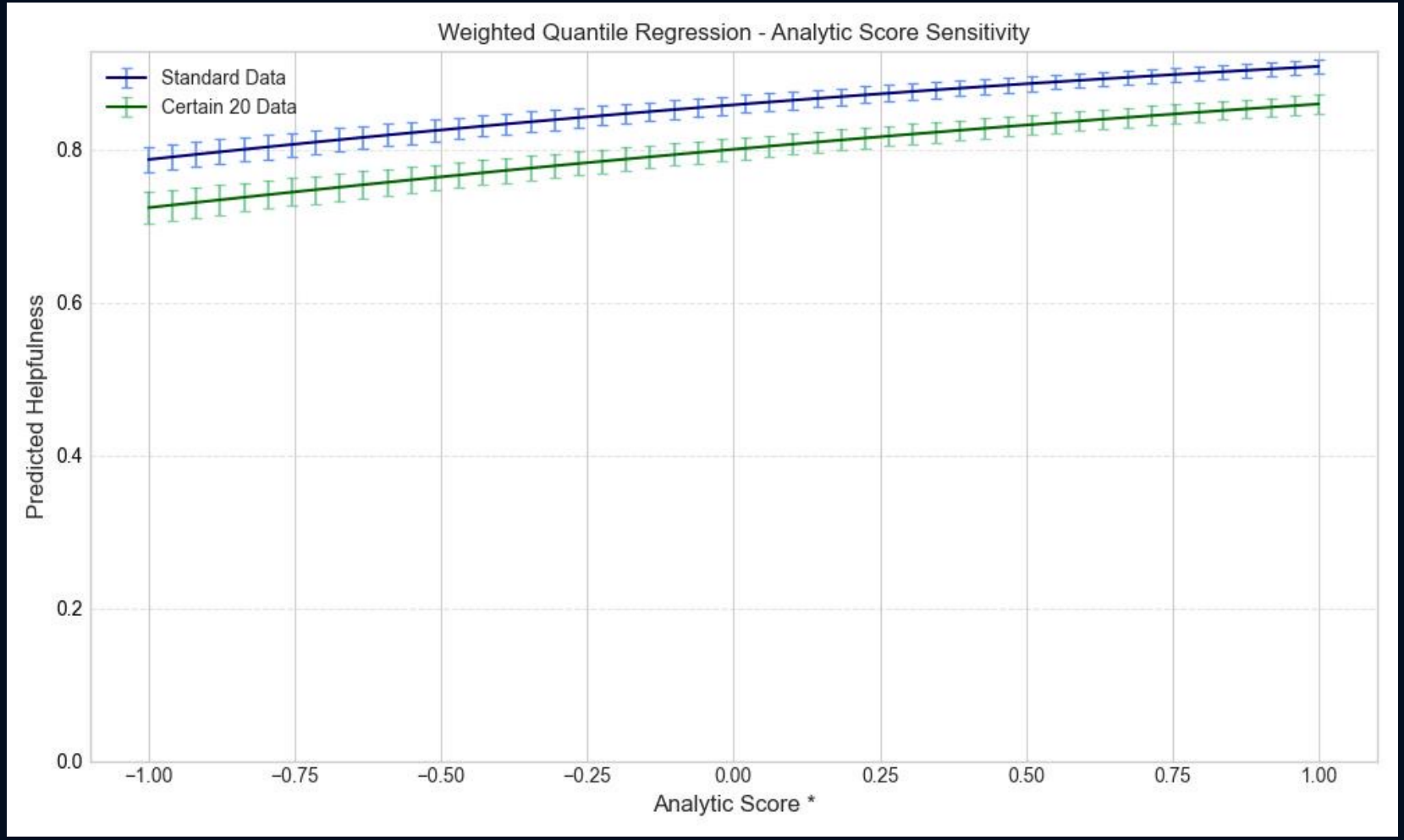


EDA: READABILITY



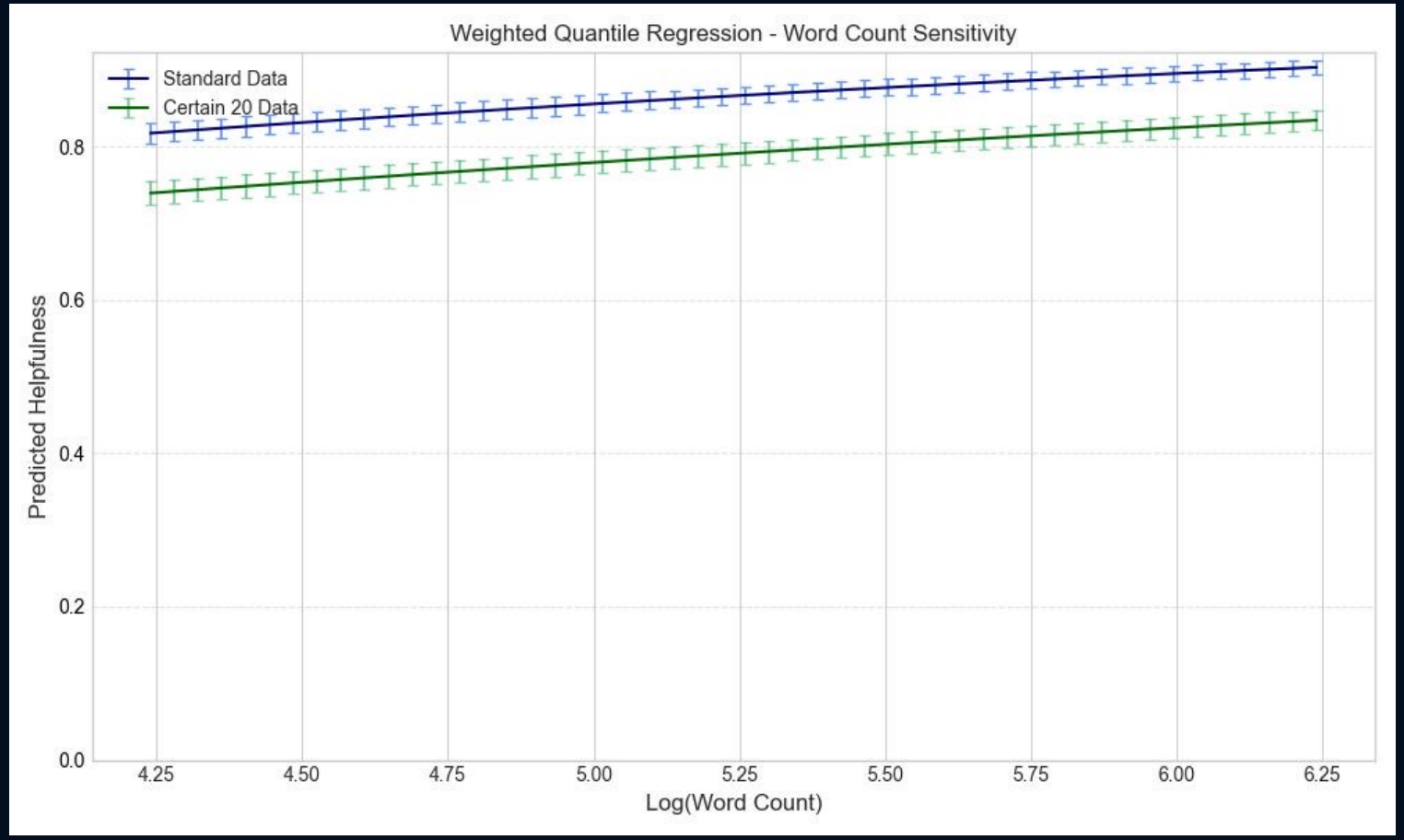
QUANTILE REGRESSION

SENSITIVITY ANALYSIS



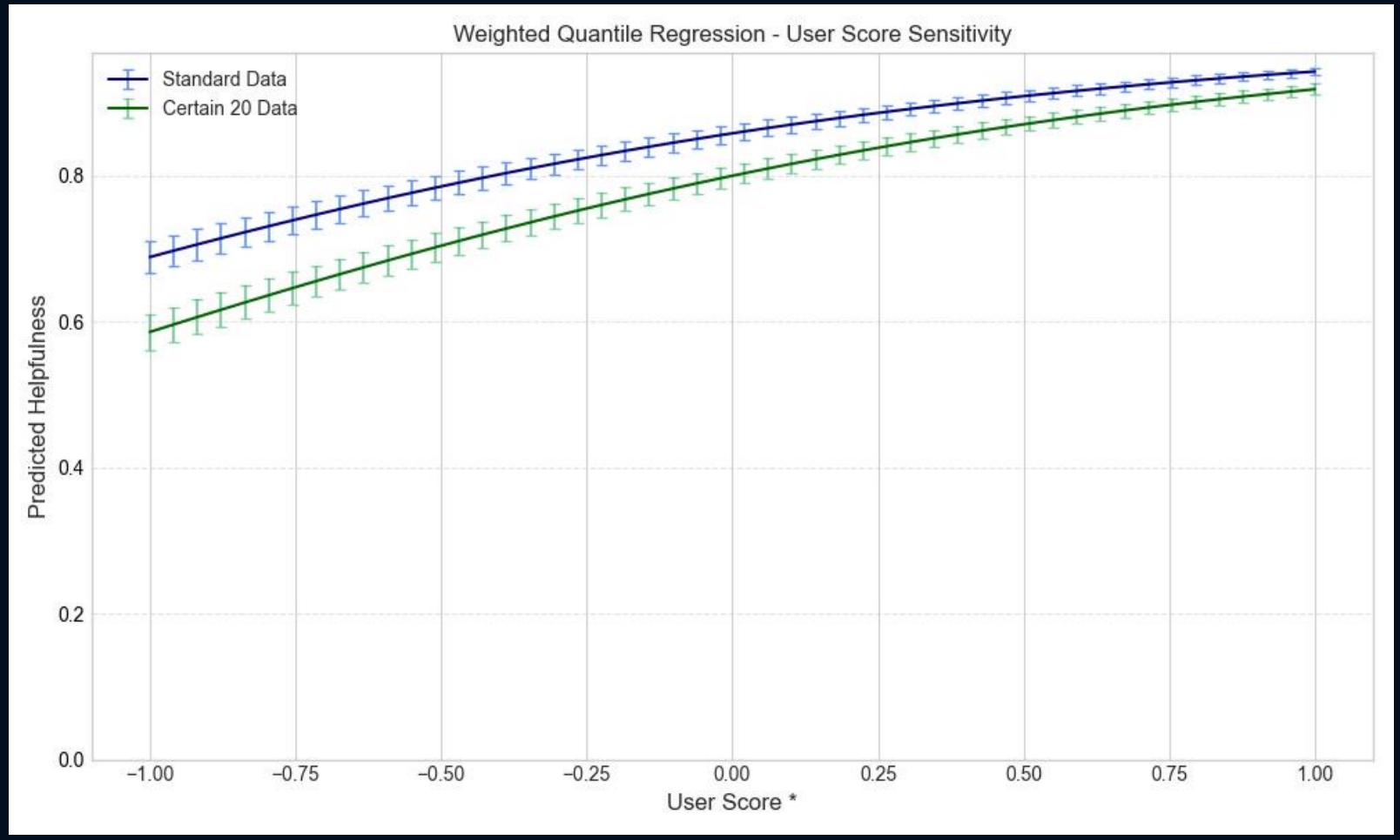
QUANTILE REGRESSION

SENSITIVITY ANALYSIS



QUANTILE REGRESSION

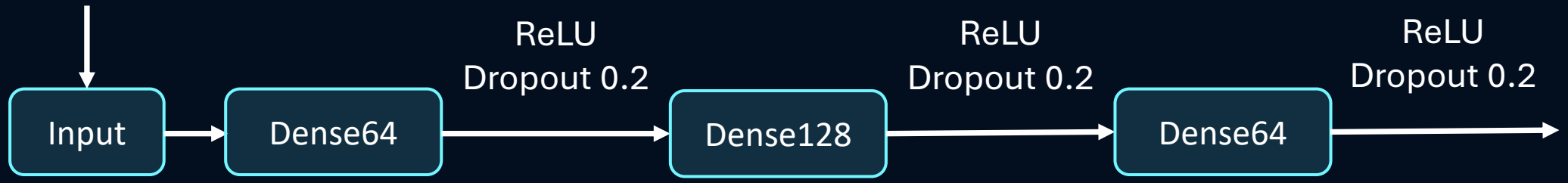
SENSITIVITY ANALYSIS



NEURAL NETWORK

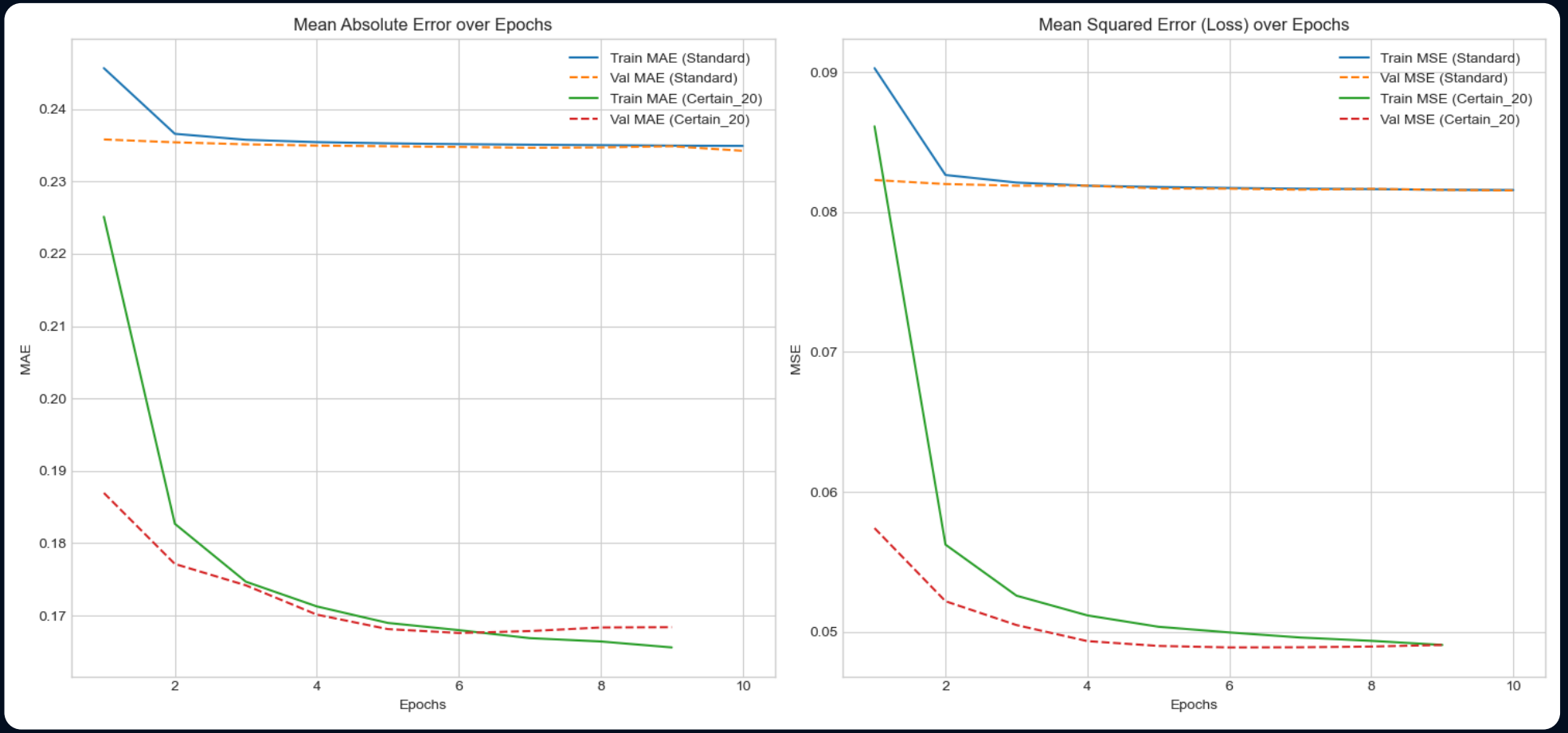
METHOD

```
review/length
review/score
sentiment_score
```



NEURAL NETWORK

RESULTS



NEURAL NETWORK

RESULTS