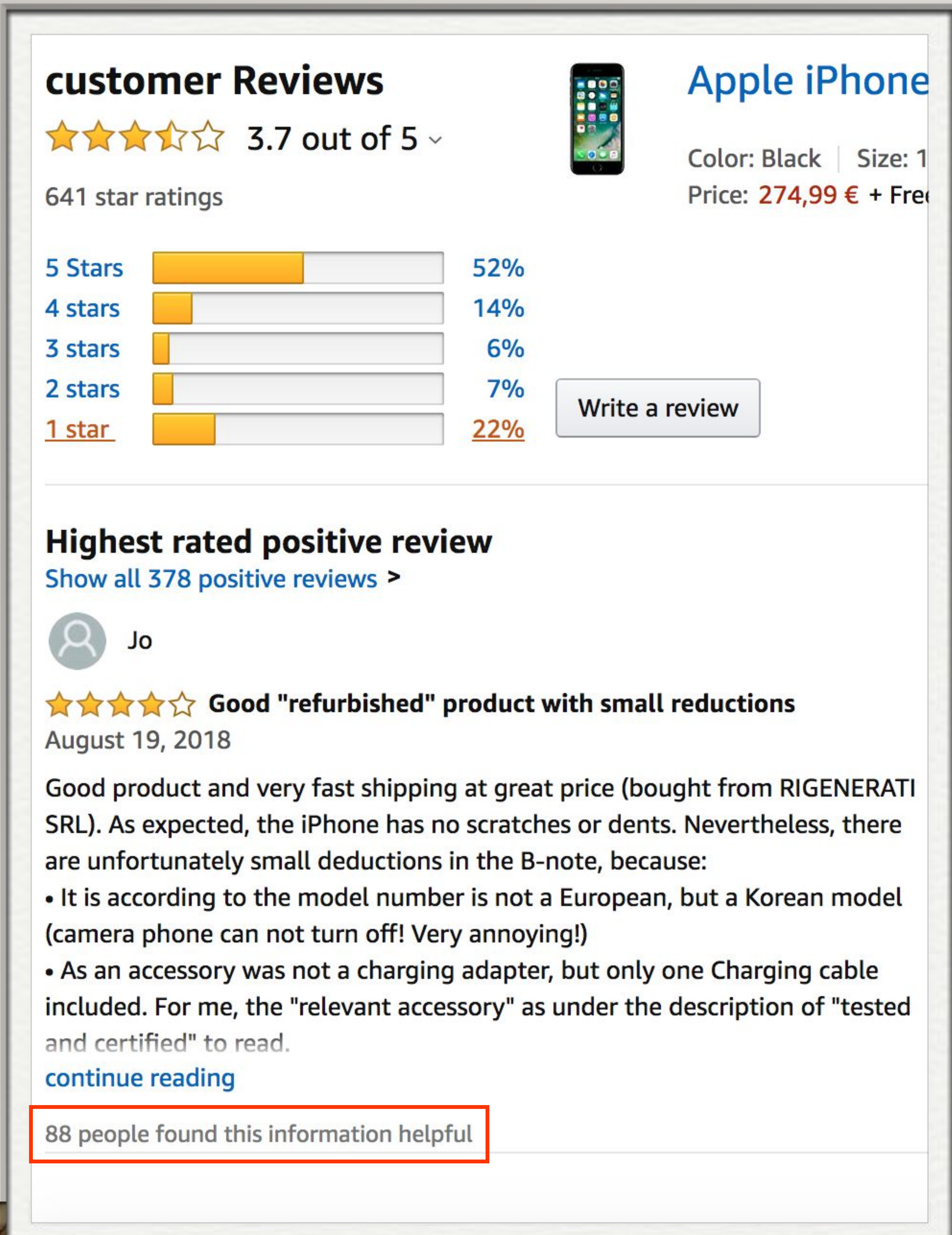
COMMENTER-BASED PREDICTION ON THE HELPFULNESS OF ONLINE PRODUCT REVIEWS

MELIH & YUJI



INTRODUCTION

WHAT IS HELPFULNESS

An indicator of a useful review about a product

WHY TO PREDICT HELPFULNESS

- Improve user experience
  - Minimize the time spend on shopping
- Increase the sales
  - The most preferred site to buy
  - More time to search for other products

FROM THE PRESS

…Because of a very subtle yet clever feature, Amazon makes the best of both the positive and negative reviews easy to find. And that feature, based on our calculations, is responsible for more than $2,700,000,000 of new revenue for Amazon every year. Not bad for what is essentially a simple question: ”Was this review helpful to you?“

- UX Articles By UIE

LITERATURE REVIEW

- Kim et al. (2007)
  - an **automatic** approach to assess the helpfulness
  - experiment with various features, such as **structural** and **lexical** features
- O'Mahony et al.(2010)
  - introducing **readability** measures on classification
- Martin et al. (2014)
  - extracting the **emotionality** from the review text by introduced a new feature GALC

LITERATURE REVIEW

- Yang et al. (2015)
  - introducing **semantic** features such as LIWC and INQUIRER
- Yang et al. (2016)
  - exploring **aspect-based** features
- Liu et al. (2016)
  - exploring **argument-based** features

(The last two works above showed that intrinsic features can support better cross-domain generalization and provide some insights and explanations for the prediction results.)

KEY ETHICAL AND SUSTAINABILITY ISSUES

1. Higher levels of economic productivity (match SDG 8.2)
2. Improved efficiency of computer resource and energy use (match SDG 9)
3. Increased added value of goods (as technology development, research and innovation) (match SDG 9.b)

RESEARCH QUESTION, HYPOTHESIS, GOAL

Problem: Will commenter-based features improve the prediction task?

Hypothesis: Helpfulness is strongly related to the commenter of one review.

Goal: Ranking the reviews on their predicted helpfulness level.

METHODS - BASELINES

Previously used features in the related field as baselines.

- **STR (Structure)** (Kim et al., 2006; Xiong and Litman, 2011)
  - total number of tokens
  - total number of sentences
  - average length of sentences
  - number of exclamation marks
  - the percentage of question sentences
- **UGR (Unigram)** (Kim et al., 2006; Xiong and Litman, 2011; Agarwal et al., 2011)
  - **Very reliable features**
  - **M:** a v**ocabu**lary with all stop-words and non-frequent words (df < 3) removed
  - **M:** the size of the vocabulary was limited to 1000
  - **M:** each review is represented by the vocabulary with TF-IDF weighting for each appeared term

METHODS - BASELINES

- **GALC (Geneva Affect Label Coder)** (Scherer, 2005)
    - proposes to recognize 36 effective states commonly distinguished by words.
    - **M:** construct a feature vector with the number of occurrences of each emotion plus one additional dimension for non-emotional words
- **SEN (Semantic features)** (Yang et al., 2015)
    - are introduced to describe the sentiment in texts
    - LIWC and General Inquirer (INQUIRER) were used

  M: since we don't have free access to the dictionary data, we extracted and designed two new semantic features:

    - sentiment polarity [-1, +1]
    - subjectivity [0, +1]

  by open source Natural Language Processing tool TextBlob

METHOD - PROPOSED

- **USR (Commenter Based Features)** (Melih & Yuji, 2019)
- Extract commenter-based features by focusing on statistics computed on the users' historical information.

  Example:

ReviewerID	Time	Overall	Sentiment Polarity	⋯	USR_Overall	USR_Sentiment_Polarity
Lisa	2018-09-30	4	0.18	⋯	nan	nan
Lisa	2018-12-07	3	0.8	⋯	4	0.18
Lisa	2019-04-25	1	0.42	⋯	(4+3)/2=3.5	(0.18+0.8)/2=0.49
Lisa	2019-09-19	4	0.3	⋯	(4+3+1)/3=2.67	(0.18+0.8+0.42)/3=0.47
Lisa	2019-11-10	5	0.94	⋯	(4+3+1+4)/4=3	(0.18+0.8+0.42+0.3)/4=0.43

- Extract commenter-based features by FunkSVD

$$H_{m\times n} = U_{m\times k}^{T} \times P_{k\times n}$$

where H is the helpfulness matrix, U could represent user information, and P could represent product information. We use training data to obtain those information.

DATASET

Dataset: Amazon Review Dataset (*Amazon.com*)

Size of Dataset: ~ 90 million reviews between 1995 and 2013.

Used subset: A subset of 37000+ reviews from category *Amazon Instant Video*.

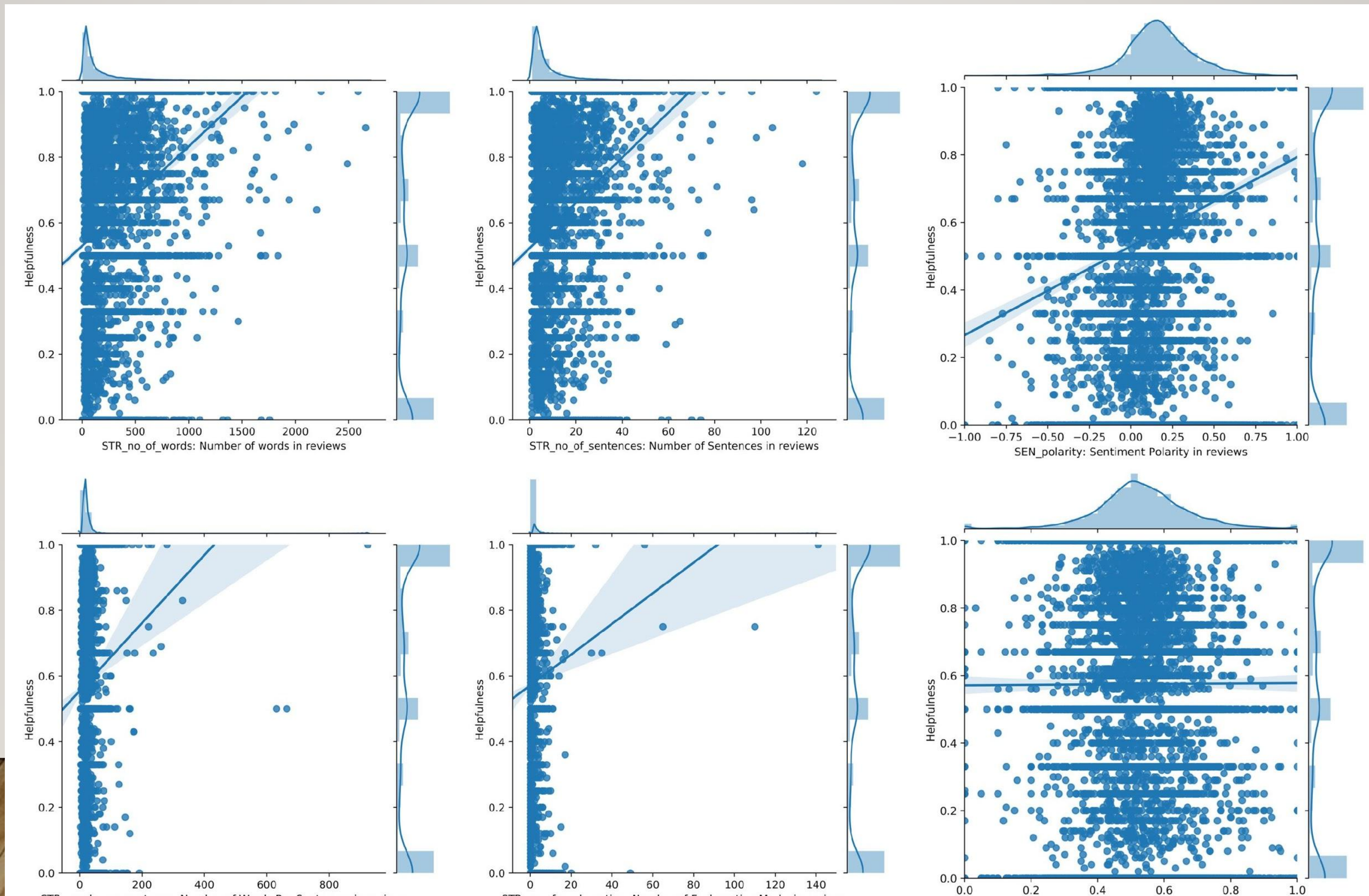
Filtered: 13000+ reviews with at least 5 votes

Train/Test Size: 9073/4060, where train set consists of data before 2014-01-01


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

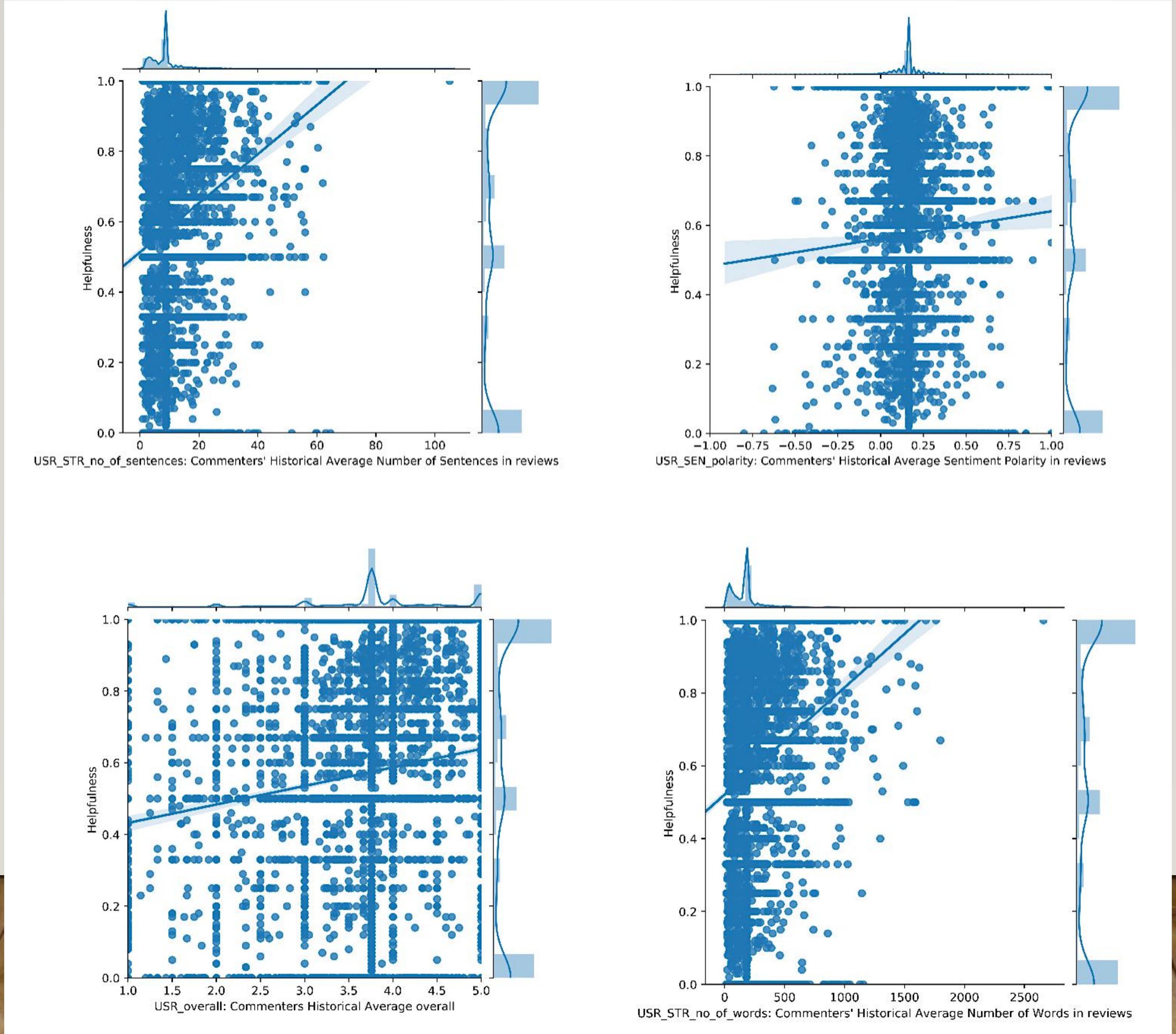
EXPLORATORY DATA ANALYSIS

Show how STR and SEN features could influence the helpfulness.



EXPLORATORY DATA ANALYSIS

Show how USR features could influence the helpfulness.



RESULTS AND ANALYSIS

Horizontal comparison - Different models but the same features(All)

Regressors	MSE	MAE
Linear Regression	0.15892	0.33997
Decision Tree	0.26599	0.38514
Gradient Boosting	**0.14378**	0.33055
LightGBM	0.14651	**0.33042**
AdaBoost	0.14717	0.33526
XGBoost	0.14387	0.33080
Random Forest	0.15738	0.33427
SVM	0.15693	0.33462

Compared to SVM, which was widely used by researchers before, boosting algorithms performed better and ran faster.

RESULTS AND ANALYSIS

Longitudinal comparison - Different features but the same model (GBDT)

Feature Sets	MSE	MAE
STR	0.16380	0.35664
UGR	0.15768	0.35050
STR + UGR	0.15706	0.34943
STR + UGR + GLAC + SEN (Baseline)	0.14417	0.33113
USR + STR + UGR + GLAC + SEN (USR + Baseline)	**0.14378**	**0.33055**

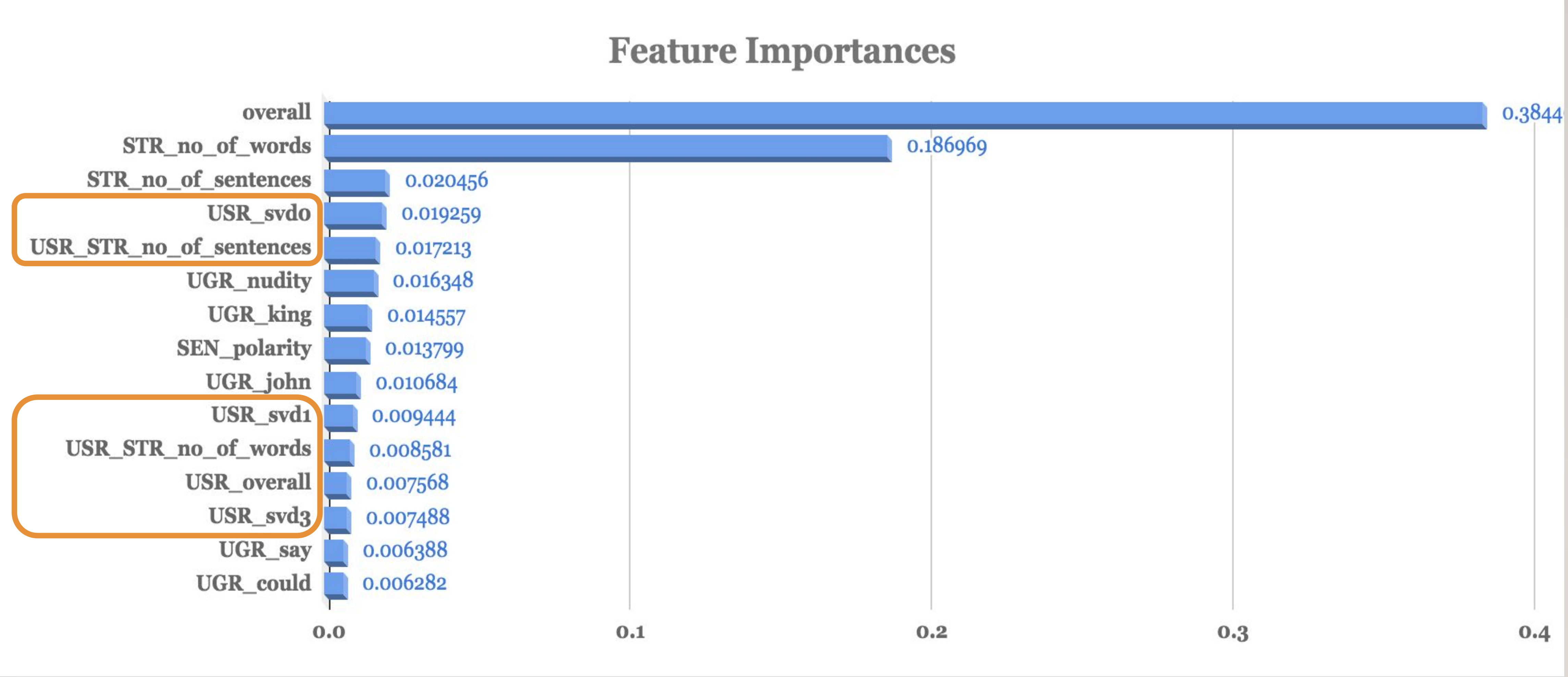
Table 1 Results on different feature sets

Feature Sets	STR	UGR	STR + UGR	Baseline	USR + Baseline
Average Rank	4	4	3.75	1.75	**1.5**

Table 2 Average rank on MSE of different features sets of 8 models

RESULTS AND ANALYSIS

Feature Importance generated from GBDT with all features:



CONCLUSION

1. Experiment results verified our hypothesis that helpfulness is related to commenters and users information could improve the prediction task
2. Boosting algorithms are better than SVM
3. Cold Start Problem: only 69.3% of the commenters have more than one comment