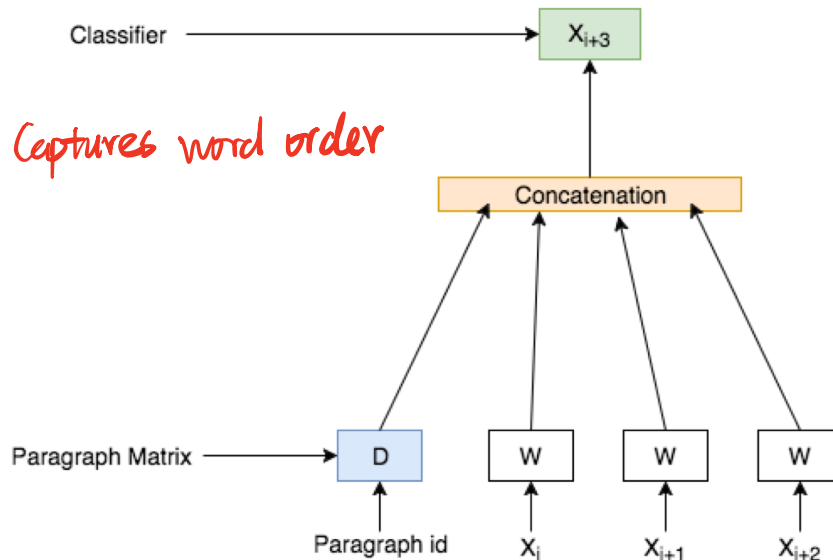# 1. Data pre-processing (Data cleaning):

1. Removing hyperlinks.
2. Removing unwanted spaces between words.
3. Converting informal words such as 'I'll', 'I've' to its formal form 'I will', 'I have', etc.
4. Adding spaces between punctuation. For example, 'This is great!It works.' is converted to 'This is great ! It works .'. Punctuations are treated as separate tokens to try to improve the accuracy of the classifier.

2. Paragraph vectors (Two flavors):
① Distributed Memory Model (PV-DM)

Classifier ──────────────→ X_{i+3}

Captures word order

Concatenation

Paragraph Matrix ──────→ D   W   W   W

Paragraph id X_i X_{i+1} X_{i+2}

- every paragraph is mapped to column of D
- every word is mapped to column of W
- given a context sampled from a paragraph (ex. X_i, X_{i+1}, X_{i+2}). the model predictes the next word X_{i+3}.
☆ The model updates both the paragraph matrix and the word matrix while training to minimize error.

② Distributed bag of words model (PV-DBOW)

Classifier → X_i X_{i+1} X_{i+2} X_{i+3}

Doesn't consider word ordering

Paragraph matrix → D
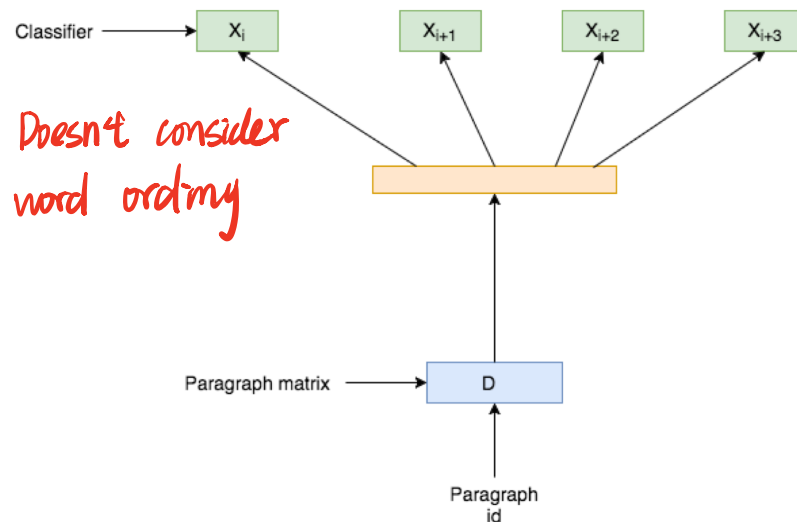
Paragraph id

- The model samples a random context from a paragraph then a random word from this context.
- Then base on that word, the model predict the context.

3. Learning product embedding:
   (using RNN with Gated Recurrent Unit (GRU))

Algorithm 3: Training GRU

    for i=1 to Number of Epoch **do**
        for Sequence S in training sequences **do**
          Train GRU with S
          if New product or user sequence starts **then**
            Reset hidden states
          end if
        end for
        Validate GRU with validation set
    end for

☆ Using GRU to combat vanishing gradient problem

1. Data Cleaning

2. Paragraph vector

300 dimensional vector

RNN

Review embedding computed using Paragraph Vector

Grouped by product and sorted by review time

GRU

Product embedding computed using GRU

128 dimensional vector

Concatenation

428 dimensional vector

SVM classification → Classify the review to "positive", "neutral" or "Negative"

Predicted Rating

Request

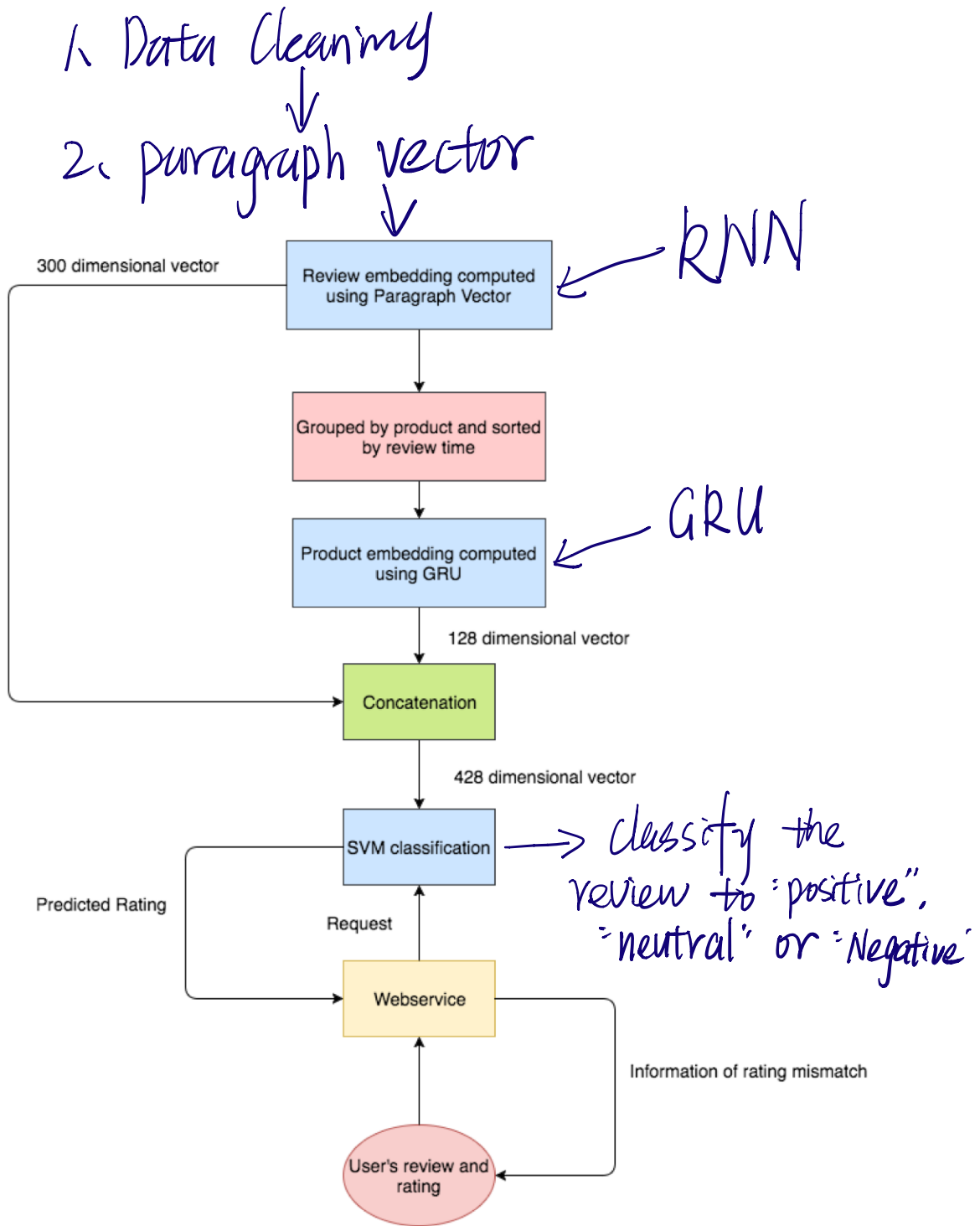Webservice

Information of rating mismatch

User's review and rating

Figure 1 Sentiment analysis of Amazon.com reviews and ratings