CASE STUDY - ECOMMERCE
TEXT MINING (CLASSIFICATION)

BUSINESS PROBLEM:

Technological advances over the past decade have led to the proliferation of consumer review websites such as Yelp.com. With the click of a button, one can now acquire information from countless other consumers about products ranging from restaurants to movies to physicians.

Yelp users give ratings and write reviews about businesses and services on Yelp. These reviews and ratings help other Yelp users to evaluate a business or a service and make a choice. While ratings are useful to convey the overall experience, they do not convey the context which led a reviewer to that experience. For example, consider a yelp review about a restaurant which has 4 stars:

"They have the best happy hours, the food is good, and service is even better. When it is winter we become regulars".

If we look at only the rating, it is difficult to guess why the user rated the restaurant as 4 stars. However, after reading the review, it is not difficult to identify that the review talks about good "food", "service" and "deals/discounts" (happy hours).

The objective of the case study is to predict yelp review start rating.

DATASET:

The data is a detailed dump of Yelp reviews, businesses, users, and checkins for the Phoenix, AZ metropolitan area. Our data contains 10,000 reviews, with the following information for each one:

1. **business_id** (ID of the business being reviewed)

2. **date** (Day the review was posted)

3. **review_id** (ID for the posted review)

4. **stars** (1–5 rating for the business)

5. **text** (Review text)

6. **type** (Type of text)

7. **user_id** (User's id)

8. {**cool / useful / funny**} (Comments on the review, given by other users)