1. Data Collection (Journal Data, Tweet, facebook posting, customer comments etc)
2. Data Cleaning (preprocessing)

Eg. Filtering out stop words,

stemming,

tokenize...

1. Build the top words model:

Proposed steps:

Input: (a list of processed text)

1. Get the absolute frequency of each words

**assigned polarity score for each words**

**pick 5 most frequent used positive words &negative words (ignore neutral words)**

1. Get the IF-IDF frequency of each words

(advs: this score considering the number of days that user used each word. If one user mentioned a specific event many times in one day, this score can smooth this bias.

**assigned polarity score for each words**

**pick 5 most frequent used positive words &negative words (ignore neutral words)**

![A screenshot of a cell phone

Description automatically generated]()