UNSUPERVISED LEXICON-BASED SENTIMENT ANALYSIS & TOPIC MODELING ON YELP REVIEW DATASET

Presented by Yan Yu Chen, Qisi Deng, Hengxin Li, Bochao Zhang

UNSUPERVISED LEXICON-BASED SENTIMENT & TOPIC MODELING (ULSTM)

Assesses the sentiment of a Yelp review by assigning a sentiment score based on semantic meaning and syntactic structure of the review, followed by Latent Dirichlet Allocation to extract global topics as highlights and/or opportunities for improvement of the restaurants.
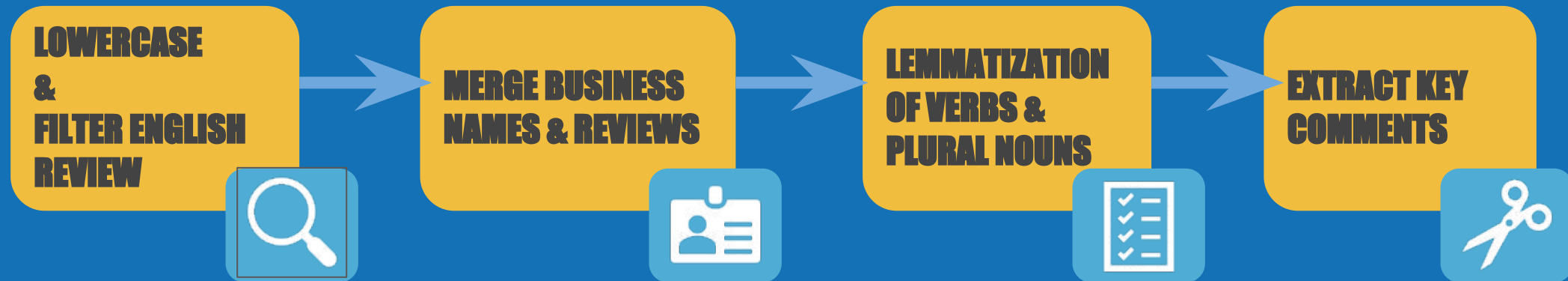
- Lexicon-Based Sentiment Model
- Sentiment Score
- Topic Modeling based on score categories

UNSUPERVISED SENTIMENT

The model is essentially an _unsupervised_ sentiment model, as it does not require any prior knowledge or label on the data of interest, which is not readily available for all types of sentiment task. Moreover, people's quantifications of the same sentiment are highly subjective, and it's not feasible to standardize these sentiment measures, which makes unsupervised sentiment analysis a superior model as the sentiment score is entirely based on the semantic meaning and syntactic structure of the review.

★★★★ ☺ ← 'The Sushi was OK.' → ☹ ★★★

PREPROCESSING

- LOWERCASE & FILTER ENGLISH REVIEW
- MERGE BUSINESS NAMES & REVIEWS
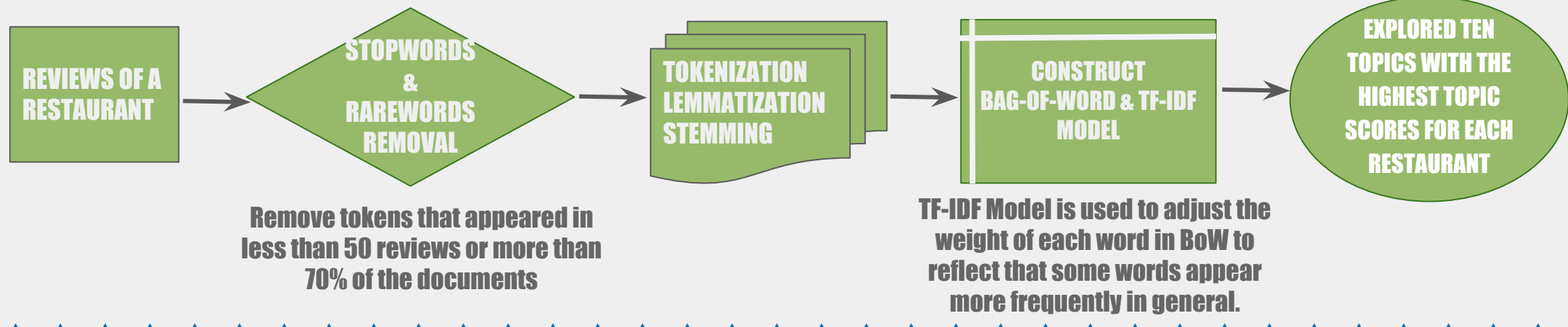- LEMMATIZATION OF VERBS & PLURAL NOUNS
- EXTRACT KEY COMMENTS

LATENT DIRICHLET ALLOCATION

- Categorized them into three sentiment levels (positive, neutral, negative) based on magnitude of sentiment scores.
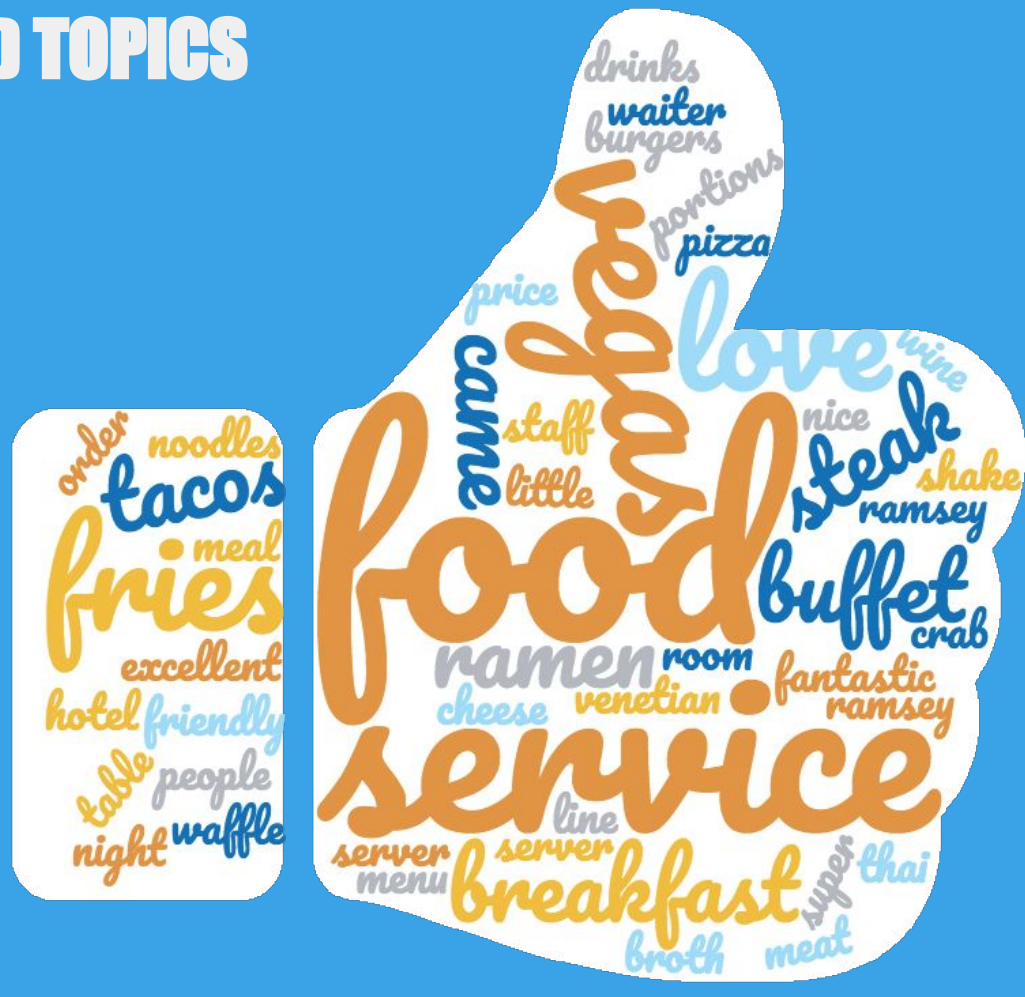- Applies LDA on reviews from the same restaurant to extract latent topics of each review and obtained the most frequently mentioned topics.
- Each topic can be treated as an aspect of dining experience in that specific restaurant (taste, price, service, environment, etc)

These aspects reflect areas that a restaurant needs to maintain the quality of or make immediate improvements on.

REVIEWS OF A RESTAURANT → STOPWORDS & RAREWORDS REMOVAL → TOKENIZATION LEMMATIZATION STEMMING → CONSTRUCT BAG-OF-WORD & TF-IDF MODEL → EXPLORED TEN TOPICS WITH THE HIGHEST TOPIC SCORES FOR EACH RESTAURANT

Remove tokens that appeared in less than 50 reviews or more than 70% of the documents

TF-IDF Model is used to adjust the weight of each word in BoW to reflect that some words appear more frequently in general.

SELECTED TOPICS

FUTURE IMPROVEMENTS OF 'WickedSpoon'

HIGHLIGHTS OF 'SkinnyFATS'

ESTABLISH SENTIMENT LEXICON

We developed a sentiment lexicon more pertinent to a restaurant setting based on original Yelp Review dataset.

- **DEFINE SEED WORDS**
  - Descriptive words with easily distinguishable polarities.
  - Positive, Negative, Neutral.
- **DEFINE LAYER & PARENT**
  - Layer is to track the position of a specific word in lexicon.
  - Parent is to identify where the word has been stemmed from.
- **GROW WORD TREES**
  - Find synonyms of 'parent' to grow dictionary to the next layer.

Assumptions:
1. Synonyms of a word share identical semantic orientation with parent.
2. When a word can be stemmed from a positive word AND a negative word, sentiment will be prioritized to the one that appears at an earlier layer.
3. When a word simultaneously occurs under positive (or negative) and neutral, then sentiment will be prioritized to non-neutral one.
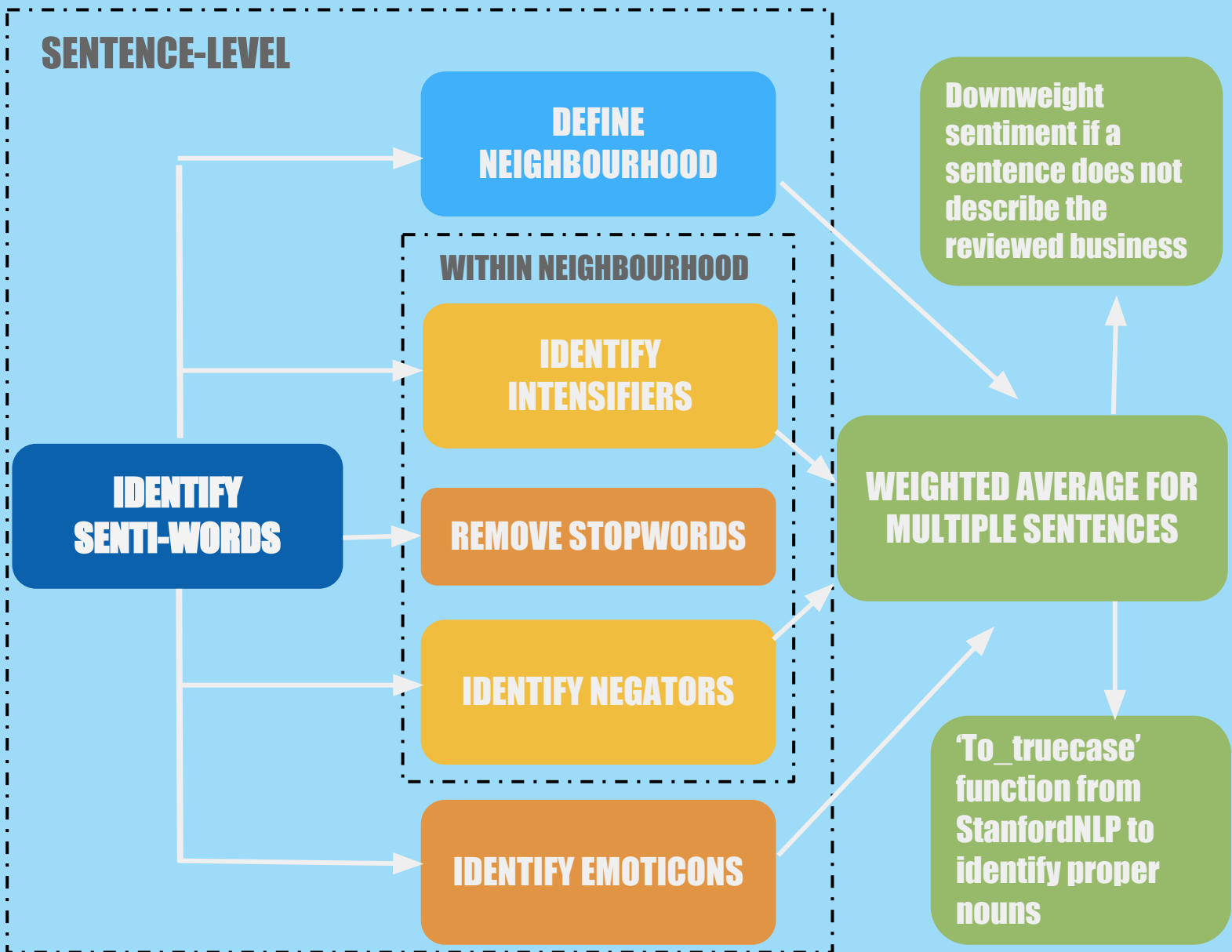
SENTIMENT SCORE EXAMPLES

REVIEWS	SENTIMENT SCORE	LABEL	REVIEW STARS
The service is terrible and while walking to the bathroom I got a glimpse off the kitchen it was so dirty and disgusting how the health department has not shut them down yet I have no idea I know for a fact i will never there again.	-4	NEGATIVE	★
Excellent food+ excellent service, everyone is very nice and helpful especially our nice and friendly host Constantine! We will definitely come back again next time we are in Vegas.	5.825	POSITIVE	★★★★★
Makis are good but globally meals are over-priced \& hyped. Go off-strip if you want to have Japanese for half the price!	0.5	NEUTRAL	★★★

SENTIMENT CALCULATOR

- Numerical Expressions:
  Positive (+1), Negative (-1), Neutral (0)

- Neighbourhood:
  Words in the independent clause of the 'Senti-word' with a max distance of 7 from 'Senti-Word'.

- Intensifiers and negators affect ONLY the 'Senti-Word' within neighbourhood.

- Emoticons are treated as special 'Senti-Words' which won't be affected by intensifiers and negations.

An example with sentiment score = (1.5)*1 + 2 = 3.5:
(x 1.5) (+1)
The food here tastes very good | and fresh :)
(+1)

SENTENCE-LEVEL

IDENTIFY SENTI-WORDS → DEFINE NEIGHBOURHOOD

WITHIN NEIGHBOURHOOD
- IDENTIFY INTENSIFIERS
- REMOVE STOPWORDS
- IDENTIFY NEGATORS
- IDENTIFY EMOTICONS

Downweight sentiment if a sentence does not describe the reviewed business

WEIGHTED AVERAGE FOR MULTIPLE SENTENCES

'To_truecase' function from StanfordNLP to identify proper nouns

EXAMPLES OF SENTI-WORD TREES

['friendly', 'Positive']

['sketchy', 'Negative']