Don't talk, Emoji it 😃🕵️📙

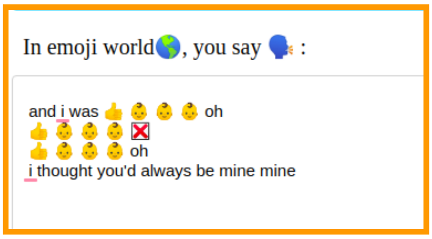
Motivation

Emoji is one common language used by people across the world from different language backgrounds. The Emoji meaning is often richer than a single word in English which makes it easy to read but hard to write. Hence, we need help from Machine Learning.

Objective

Provide a solution to translate sentences from English to Emojis.

Sample output for Justin Bieber Song "Baby" Lyric



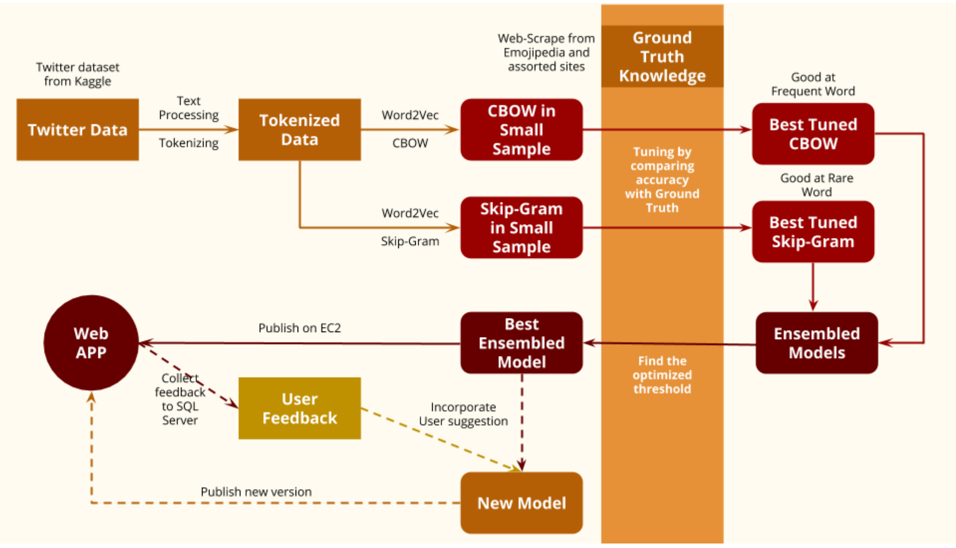
In emoji world🌎, you say 🗣 :

and i was 👍 👶 👶 👶 oh
👍 👶 👶 👶 ❌
👍 👶 👶 👶 oh
i thought you'd always be mine mine

Methodology

- **Data Collection:**
  - **19 million** Twitter records contain at least one Emoji
  - **Web Scraping** Emoji "true meaning" from various Emoji website as Ground Truth knowledge
- **Text Processing:**
  - Lower all cases
  - Remove punctuation
  - Try remove stop words
  - Try Stemmer and Lemmetizer
- **Model building:**
  - Processed text into a **two-layer Neural Network** called Word2Vec
  - Tuned model hyperparameters in a small sample
  - Tuned model with text processing techniques in a small sample
  - Built 2 Word2Vec algorithms. CBOW and skip-gram have different advantages in predicting common or rare words.
  - **Ensembled** two algorithms with a threshold to determine which word is frequent.
  - The final Model recorded 62% accuracy in predicting the Ground Truth, **improved 55% compared to baseline**(7%).
- **Web APP:**
  - Built a real-time translation **Web App** with Dash
  - Published on an **AWS EC2** instance
  - Collecting user feedback into **SQL** database which enable us to adjust the model regularly

Chart: Project Workflow



Key findings

- Not limited to the original meaning, people use Emoji creatively, for instance, the emoji 🍆 is often used in flirting.
- Tuning in the text processing steps is as crucial as tuning model hyperparameter in this NLP analysis.
- The ensemble method is a solution to make bad predictors into a better one when advanced methods are not applicable. (RNN text generator requires TB level memory since we had 3000 more characters which usually 26+10)

Tech Stack

aws BeautifulSoup Flask PostgreSQL GENSIM topic modelling for humans