Sentiment Analysis of Customer Reviews

Aqsa Mahmood

Agenda

- Motivation
- Objective
- Dataset Overview
- Methodology
- Results
- Limitations and Improvements
- Conclusion

Motivation

- A huge amount of comments, opinions and reviews are shared on social medial and online resources everyday.
- Most of these reviews are unstructured which makes it hard for businesses to analyse customers feedback at scale.
- Being able to quickly determine the sentiment of these reviews helps businesses to identify customers views towards their product and services.
- This allows organizations to make intelligent decisions, improve their products and services.

Objective

“This project intends to build a binary classification model (positive/ negative) for customer reviews.”

Dataset Overview

- The Sentiment Labelled Sentences dataset was obtained from the UCI Machine Learning repository.
- The dataset is composed of 3000 text sentences extracted from reviews of products, movies and restaurants.
- Each sentence is labelled positive or negative.
- The reviews came from amazon, IMDb and yelp websites with equal ratio of positive and negative reviews.

Reviews	Sentiment
i bought it for my mother and she had a problem with the battery.	0
needless to say, i wasted my money.	0
The mic is great.	1
very good quality though	1

Dataset Overview

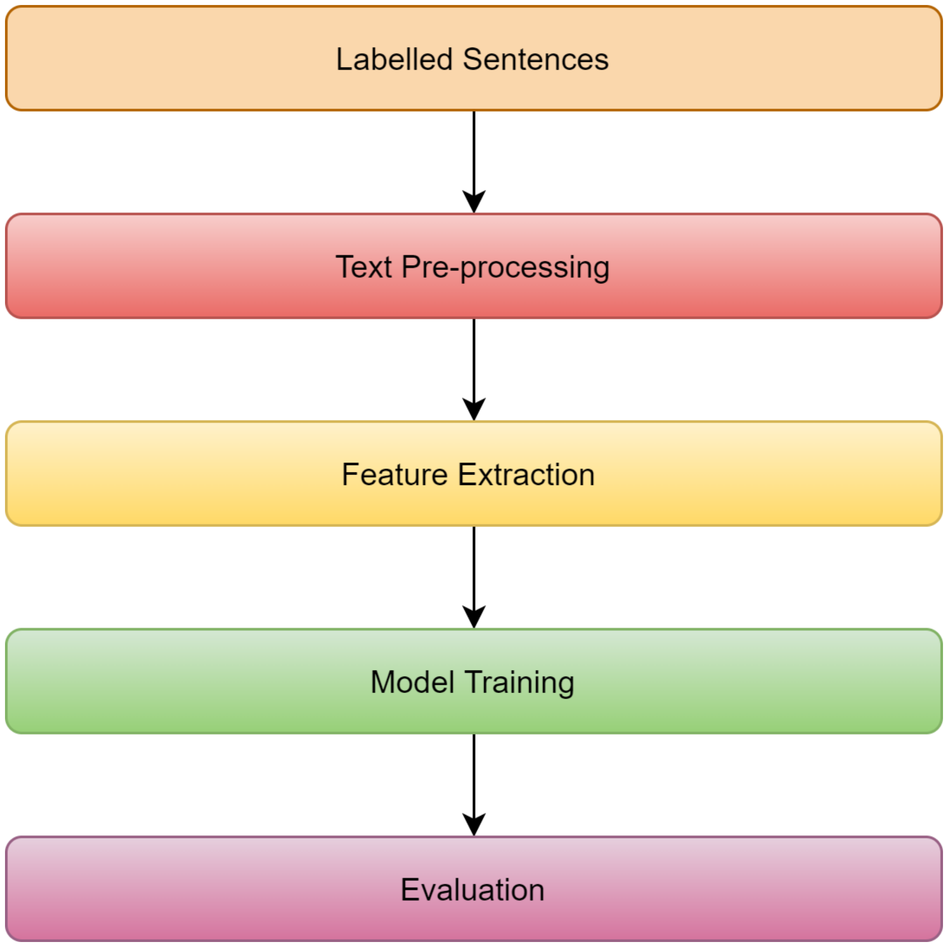
Positive Reviews

Negative Reviews

Dataset Overview

Reviews with Highest Polarity	Reviews with Lowest Polarity
Excellent bluetooth headset.	Worst customer service.
EXCELLENT SERVICE!!!!!!!!.	It's A PIECE OF CRAP!
best bluetooth on the market.	The movie is terribly boring in places.
I am also very happy with the price.	Food quality has been horrible.
The reception is excellent!	The kids play area is NASTY!

Methodology



Methodology: Data Cleaning

- Converting text to lowercase
- Removing stop words
- Removing special characters and numbers
- Tokenization
- Lemmatization

Methodology: Feature Extraction

- **Bag of words**

- Captures presence of words within text data

- **TF-IDF**

- Captures how frequently a word occurred in a document and the entire corpus.

- **Word Embeddings**

- Vector representation of words

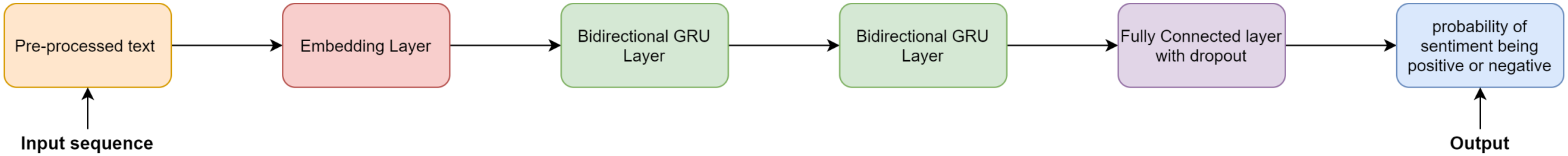
- Pretrained GloVe word embedding of 100 dimension trained using 6 billion words

- Contains 400,000 words in its vocabulary.

Methodology: Machine Learning Models

- The pre-processed dataset is divided into train (70%) and test sets (30%).
- Bag of words and TF-IDF techniques were used to extract features from text.
- The train dataset is used for training, validating and hyperparameter tuning model parameters.
- Grid search with 10 fold cross validation was used to find optimal model parameters.
- Two Machine learning models Support vector machine and Logistic regression were trained using BOW and TF-IDF features.

Methodology: Deep Learning Model



- Pretrained GloVe embeddings were used to get vector representation of words.
- Trained Bidirectional GRU model using generated word embeddings to determine sentiment of reviews.
- Adam optimizer was used to train the model and binary cross entropy with logits function was used to calculate error.
- The model was trained for 50 epochs.

Results

Model	Training Accuracy	Test Accuracy	Precision	Recall	F1 score
Logistic Regression with BOW	90%	77%	74%	77%	76%
SVM with BOW	96%	79%	78%	78%	78%
Logistic Regression with TFIDF	88%	75%	87%	69%	77%
SVM with TFIDF	94%	78%	78%	76%	77%
Gated Recurrent Unit(GRU)	98%	85%	87%	83%	85%

Limitations and Improvements

- **Limitations:**

-Grammatical errors

-Spelling mistakes

-Sarcasm in hate reviews is an unsolved problem.

- **Improvements**

-Implementing this framework on large dataset of reviews

-Correcting Spelling mistakes and excluding non vocabulary words

-Considering learning word embeddings form scratch

Conclusion

- We Classified text by sentiment using different feature extraction, Machine learning and Deep learning techniques.
- Word Embedding is the most powerful feature extraction technique as compare to Bag of words and TFIDF.
- Gated Recurrent Unit model outperforms SVM and Logistic Regression.

Thank You for Listening