Domain background

Sentiment analysis is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.

Since customers express their thoughts and feelings more openly than ever before, sentiment analysis is becoming an essential tool to monitor and understand that sentiment. Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customers' needs.

For example, using sentiment analysis to automatically analyze 4,000+ reviews about your product could help you discover if customers are happy about your pricing plans and customer service. Maybe you want to gauge brand sentiment on social media, in real time and over time, so you can detect disgruntled customers immediately and respond as soon as possible. (1)

The applications of SA are endless.

Problem Statement

The goal is to create a model from this(3) dataset and apply the result model to my company www.zooplus.com
The tasks to achieve that are the following:

1. Download and preprocess data from Amazon ***pet suppliers reviews***
2. Train a classifier that can determine if a text review feeling is good or bad
3. Make the classifier run on AWS as an endpoint

Datasets and Inputs

The dataset that is going to be used comes from (3)

and in particular from the Pet supplies package that in this moment of time has 157.836 reviews

each row has the following format


```
{
      "reviewerID": "A2SUAM1J3GNN3B",
      "asin": "0000013714",
      "reviewerName": "J. McDonald",
      "helpful": [2, 3],
      "reviewText": "I bought this for my husband who plays the piano. He is
      having a wonderful time playing these old hymns. The music is at times
      hard to read because we think the book was published for singing from more
      than playing from. Great purchase though!",
      "overall": 5.0,
      "summary": "Heavenly Highway Hymns",
      "unixReviewTime": 1252800000,
```

```
    "reviewTime": "09 13, 2009"
}
```

Where

`reviewerID` - ID of the reviewer, e.g. A2SUAM1J3GNN3B

- `asin` - ID of the product, e.g. 0000013714
- `reviewerName` - name of the reviewer
- `helpful` - helpfulness rating of the review, e.g. 2/3
- `reviewText` - text of the review
- `overall` - rating of the product
- `summary` - summary of the review
- `unixReviewTime` - time of the review (unix time)
- `reviewTime` - time of the review (raw)

Solution Statement

I will create a model based on RNN in pytorch using Sagemaker. The steps will be the following:

- Clean and process data
- Split in train and test sets
- Apply training to the model
- Hyperparam tunning
- Validate results(4)

Benchmark Model

Since the goal of the project will be to create a RNN the idea is to compare LSTM and GRU approaches.

1. The GRU controls the flow of information like the LSTM unit, but without having to use a ***memory unit***. It just exposes the full hidden content without any control.

- GRU is relatively new, and from my perspective, the performance is on par with LSTM, but computationally ***more efficient*** (*less complex structure as pointed out*). (5)

So I will measure memory, speed model creation and performance

Evaluation Metrics

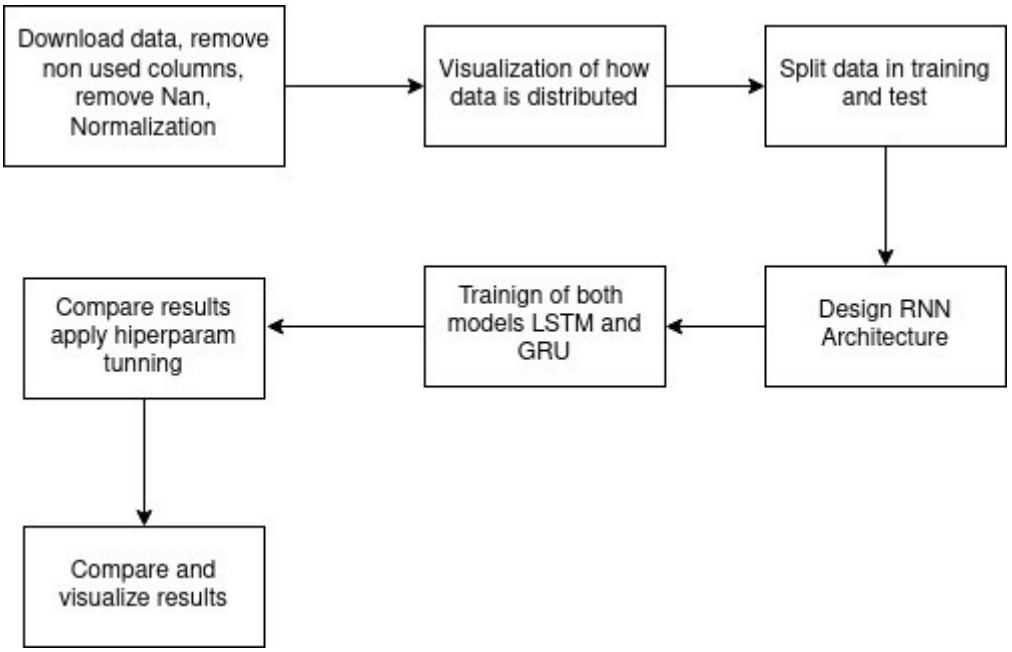
As a common metric I will use:

$$accuracy = \frac{(true_{positives} * true_{negatives})}{datasetsize}$$

A true positive when the algorithm hits a positive review
A true negative review performed hits a negative review

Project Design

The workflow for the solution will be:



- For data processing I will use Pandas
- For visualization I will use Plotly
- For splitting data Sklearn
- For RNN design Pytorch

Links

1. https://monkeylearn.com/sentiment-analysis/
2. https://analyticsindiamag.com/10-popular-datasets-for-sentiment-analysis/
3. http://jmcauley.ucsd.edu/data/amazon/
4. http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/
5. https://medium.com/analytics-vidhya/rnn-vs-gru-vs-lstm-863b0b7b1573