PRN - 18030142043

Name - Subodh Dharmadhikari

Course - Text Analytics

Class - MSc CA Data Science

Sentiment and Context Analysis using Machine Learning and Deep Learning

Amazon reviews and IMDB case study and Project

Online shopping has been growing from 10 years and many companies like Amazon,Flipkart etc have been serving to meet the increasing demand.

The merchants wants to know about how well received their product is in the market right now.They even want to know if new version of some product comes in the market how it is accepted by public compared to previous version.

The whole business domain of commerce is multi-billion space which very competitive as there are numerous new players have come into the market to disrupt the space in various categories.

Some products are e-commerce exclusive for example some products are Amazon exclusive.Their business depends on only on these platforms and reviews impact their business directly.

To control this segment of e-commerce giant companies include each verified review in the account to improve their product further.

Verified reviews are unique emotions and experiences.

Companies apply various models to get overall receive by the customers.

Various methods from simple Logistic regression to

Bi-direction LSTM CNN is used to get context and polarity of the customers.

The use of methods primarily depends on amount of data you have.

These review sentiment can help them to take many strategic decisions about their company like marketing plan, improvement of post sale services, research for launching new products and further other important to small things like packaging-accessories.

A specific product may be offered on numerous websites and the expenses may additionally range.

As clients generally want the quality to be exceptional for the lowest charge however can't at once test it but evaluations from different users seem to be the most reliable way to determine whether to buy the product or not.

Therefore sentiment analysis has established crucial to understand a product's popularity.It can be done using various methods

Naive Bayesian,

Support Vector Machine,

Decision trees or

Logistic regression

Sometimes we don't have enough reviews to get good accuracy with the help of machine learning algorithm then we can get good results with

Bidirectional CNNs in deep learning way.The deep learning when effectively used give better accuracy as it can extract more features and even smaller things are taken into consideration.

The smaller things sometimes affect the business massively for example if you have bad packaging for your product then your new version of product can lead to sudden decline in demand another could be shorter battery life of product compared to another available products in the market.

Reviews can be used in following ways

Review Analysis: Bag of words model is used to categorise documents and frequency of occurrence of words is extracted for training the classifier.

Sentiment Classification: Classifies the extracted words as positive or negative.

Sometimes using these techniques gives edge over the competitors and sometimes these techniques are used to stay relevant to market techniques.When all other competitors are using theses techniques to manage and improve their customer experience it becomes mandatory to relative companies to fine tune their businesses and be relevant.

For newly launched companies it’s mandatory from first stage as their competitors might have already captured the market and have good reputation.New companies infant of these giants can not scale if they don’t consider each experience into the account.

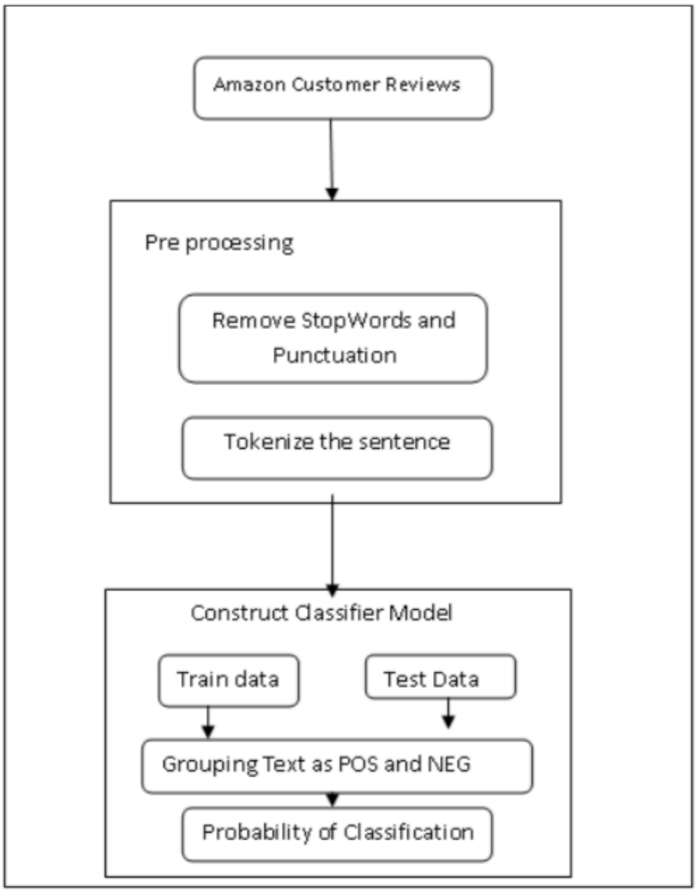
Along with sentiments of the user we have to do lot of analysis like how many reviews are coming in each month and all.That all things come under exploratory data analysis.Exploratory data analysis is nothing but visualisations of visualisation of data in the various forms like pie charts ,bar graphs etc.

These visualisations give insights to higher management to take strategic decisions and make required changes demanded by public if some older version of product has more popularity and sell they can not just discontinue it directly but they have to understand the popular features of it and keep them or incorporate in coming models as sometimes some features becomes identity of the brands like Nokia has their solid battery power as its feature and that drive certain segments of the customers towards the brand.

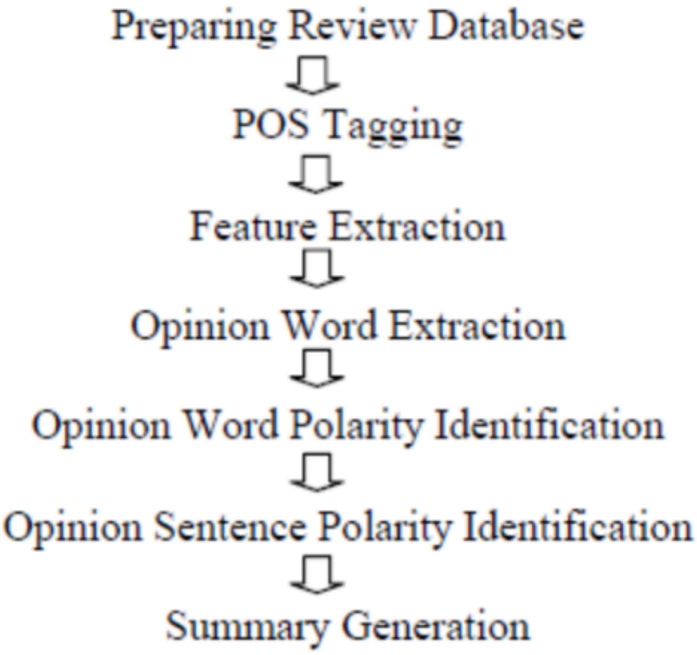
DATA SET DESCRIPTION:

- The dataset contains product reviews and metadata of 'Clothing, Shoes and Jewelry' category from Amazon, including 2.5 million reviews spanning May 1996 - July 2014.
- The dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links.

Steps for exploratory data analysis :


Amazon Customer Reviews
Pre processing
Remove StopWords and Punctuation
Tokenize the sentence
Construct Classifier Model
Train data
Test Data
Grouping Text as POS and NEG
Probability of Classification

Steps for context and polarity analysis :


Preparing Review Database
POS Tagging
Feature Extraction
Opinion Word Extraction
Opinion Word Polarity Identification
Opinion Sentence Polarity Identification
Summary Generation

Dataset

‘amazon_reviews.csv'

- Consist of all the reviews for the products in 'Clothing, Shoes and Jewelry' category from Amazon. Each review is a row in csv file.

FIELDS:

- 1 ReviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- 2 Asin - ID of the product, e.g. 0000013714
- 3 Reviewer Name - name of the reviewer
- 4 Helpful - helpfulness rating of the review, e.g. 2/3
- 5 Review Text - text of the review
- 6 Overall - rating of the product
- 7 Summary - summary of the review
- 8 Unix Review Time - time of the review (unix time)

DATA PROCESSING:

REVIEW DATA i.e. amazon_reviews.csv

- Reading a csv file and cleaning it.

- Iterating over list and loading each index as csv and getting the data from the each index and making a list of Tuples containg all the data of csv files.

- Creating a dataframe using the list of Tuples got in the previous step.

SENTIMENTAL ANALYSIS ON REVIEWS

Wordcloud of summary section of 'Positive' and 'Negative' Reviews on Amazon.

- VADER (Valence Aware Dictionary and Sentiment Reasoner) Sentiment analysis tool was used to calculate the sentiment of reviews.
- Sentiment distribution (positive, negative and neutral) across each product along with their names mapped with the product database ‘amazon_reviews.json'.

WordCloud of summary section of 'Positive' and 'Negative' Reviews on Amazon.

- Created a function to calculate sentiments using Vader Sentiment Analyzer and Naive Bayes Analyzer.
- Only taking 10000 reviews into consideration for Sentiment Analysis so that jupyter notebook dosen't crash.
- Sentiment value was calculated for each review and stored in the new column 'Sentiment_Score' of DataFrame.
- Seperated negatives and positives Sentiment_Score into different dataframes for creating a 'Wordcloud'.
- Stemming function was created for stemming of different form of words which will be used by 'create_Word_Corpus()' function. PorterStemmer from nltk.stem was used for stemming.
- Function 'create_Word_Corpus()' was created to generate a Word Corpus.

In addition to that I have tested Bidirectional LSTM

on IMDB data to check how Deep Learning works for sentiment analysis.

The Dataset is used for IMDB sentiment analysis is Stanford Data.

I have performed Bidirectional LSTM to get polarity of sentences.

Thank you.