E-commerce

Project 1

DESCRIPTION

Problem Statement

- Amazon is an online shopping website that now caters to millions of people everywhere. Over 34,000 consumer reviews for Amazon brand products like Kindle, Fire TV Stick and more are provided.
- The dataset has attributes like brand, categories, primary categories, reviews.title, reviews.text, and the sentiment. Sentiment is a categorical variable with three levels "Positive", "Negative“, and "Neutral". For a given unseen data, the sentiment needs to be predicted.
- You are required to predict Sentiment or Satisfaction of a purchase based on multiple features and review text.

Dataset Snapshot

name	brand	categories	primary Categories	reviews.text	reviews.title	Sentiment
Amazon Kindle E-Reader 6" Wifi (8th Generation...	Amazon	Computers,Electronics Features,Tablets,Electro...	Electronics	I thought it would be as big as small paper bu...	Too small	Neutral
Amazon Kindle E-Reader 6" Wifi (8th Generation...	Amazon	Computers,Electronics Features,Tablets,Electro...	Electronics	This kindle is light and easy to use especiall...	Great light reader. Easy to use at the beach	Positive
Amazon Kindle E-Reader 6" Wifi (8th Generation...	Amazon	Computers,Electronics Features,Tablets,Electro...	Electronics	Didnt know how much i'd use a kindle so went f...	Great for the price	Positive
Amazon Kindle E-Reader 6" Wifi (8th Generation...	Amazon	Computers,Electronics Features,Tablets,Electro...	Electronics	I am 100 happy with my purchase. I caught it o...	A Great Buy	Positive
Amazon Kindle E-Reader 6" Wifi (8th Generation...	Amazon	Computers,Electronics Features,Tablets,Electro...	Electronics	Solid entry level Kindle. Great for kids. Gift...	Solid entry-level Kindle. Great for kids	Positive

Project Task: Week 1

Class Imbalance Problem:

1. Perform an EDA on the dataset.
    - See what a positive, negative, and neutral review looks like
    - Check the class count for each class. It’s a class imbalance problem.
1. Convert the reviews in Tf-Idf score.
2. Run multinomial Naive Bayes classifier. Everything will be classified as positive because of the class imbalance.

Tackling Class Imbalance Problem:

1. Oversampling or undersampling can be used to tackle the class imbalance problem.
2. In case of class imbalance criteria, use the following metrices for evaluating model performance: precision, recall, F1-score, AUC-ROC curve. Use F1-Score as the evaluation criteria for this project.
3. Use Tree-based classifiers like Random Forest and XGBoost.

Note: Tree-based classifiers work on two ideologies namely, Bagging or Boosting and have fine-tuning parameter which takes care of the imbalanced class.

Project Task: Week 2

Model Selection:

1. Apply multi-class SVM's and neural nets.
2. Use possible ensemble techniques like: XGboost + oversampled_multinomial_NB.
3. Assign a score to the sentence sentiment (engineer a feature called sentiment score). Use this engineered feature in the model and check for improvements. Draw insights on the same.

Applying LSTM:

1. Use LSTM for the previous problem (use parameters of LSTM like top-word, embedding-length, Dropout, epochs, number of layers, etc.)

Hint: Another variation of LSTM, GRU (Gated Recurrent Units) can be tried as well.

1. Compare the accuracy of neural nets with traditional ML based algorithms.
2. Find the best setting of LSTM (Neural Net) and GRU that can best classify the reviews as positive, negative, and neutral.

Hint: Use techniques like Grid Search, Cross-Validation and Random Search

Topic Modeling:

1. Cluster similar reviews.

Note: Some reviews may talk about the device as a gift-option. Other reviews may be about product looks and some may highlight about its battery and performance. Try naming the clusters.

1. Perform Topic Modeling

Hint: Use scikit-learn provided Latent Dirchlette Allocation (LDA) and Non-Negative Matrix Factorization (NMF).

Download the Data sets from here **.**