黃秋月-H34075033

Project title

“Sentiment Analysis on Amazon Kindle Book Review”

Background and motivation

Nowadays, demand for e-commerce is increasing. E-shopping vendors like Amazon have grown in popularity. Amazon is one of the companies that provide reviews of users regarding all its products. Customers express their supposition or estimation by giving feedbacks within the frame of text. Sentiment analysis examination is the method of deciding the supposition or feeling communicated as either positive or negative. The voice of customer concepts such as reviews and customer experiences are getting more important since customers purchase items without seeing or touching them. Consumer comments in online forums have been demonstrated to be a valuable source for revealing consumer insights, and this user-generated content represents a promising alternative source for possibly distinguishing client needs. In this way, mining this user-generated content and analyzing the emotions of the comments communicated by customers might be useful for companies. Many researchers highlight the importance of factoring in user-generated content to aid in decision-making in the marketing field. Particularly, from online reviews consumers generate a huge amount of brand-related information that includes a decisive potential business value for marketing purposes.

This research aims to give a realistic visual analysis of customer feedback sentiment analysis and visualization approaches. In this study, Amazon book reviews were utilized to demonstrate the value of the suggested interactive dashboard in assisting decision-makers. A sentiment analysis system of textual reviews was built utilizing a huge dataset of the reviews to achieve the goal. In addition, visualization techniques are utilized to offer consumers a summary in an interactive form utilizing multiple representations.

Literature review

Sentiment analysis, also known as opinion mining, investigates, analyzes, and extracts subjective texts containing users' views, preferences, and sentiments.

Since Bo Pang in the paper titled " A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. " put forth this concept in 2002, academics have conducted a broad variety of relevant study of its usefulness in opinion monitoring and corporate competitive intelligence

Yadav, M. P., et al. used K Mean to explain customer behavior for E-commerce firms in their work titled “Mining the consumer behavior using web use mining in e-commerce.” With the rapid expansion of the internet, people may now quickly discover, extract, filter, and assess whatever they desire. With technological advancements, servers can now be selected from a superstore. Several techniques were used by the e-commerce industry to gain the trust of the customer and providing ratings about the product so that the user can recommend the product review and other details about the product that is available on the e-commerce website to his colleges, friends, and relatives.

In their work titled “Web Mining Techniques in E-Commerce Applications,” Ahmad Tasnim Siddiqui et al. highlighted that purchasing has grown as compared to window shopping since it gives millions of ranges. As a result, businesses are able to attract the majority of clients. The fundamental technique may be done to secure the data as time progresses, and hence the progression. Technology was likewise changing at a rapid pace from year to year. As a result, users have become accustomed to paying a high price to the vendor due to added costs since the nineteenth century.

Songbo Tan et al. determine the particular condition of the classifier about which the classifiers are being determined in order to find out the classification of the particular segment to find the token word from the given review by the customer in their paper titled “Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis.” The e-commerce industry is capable of providing good quality products and services such that the customer's belief in purchasing the product is recommended to others, and the e-commerce organization will be satisfying the customer's expectation to buy the particular product online on an e-commerce website.

Research method

Sentiment analysis is described as the task of determining writers' views on certain items. The views produced by thought leaders and regular people influence people's decision-making processes. When a person wants to purchase products online, he or she will usually begin by looking for reviews and comments published by other people about the various options. Today, there is a massive burst of sentiments accessible through social media platforms such as Twitter, Facebook, message boards, blogs, and user forums. These text snippets are a gold mine for businesses and people that wish to monitor their reputation and receive quick feedback on their products and performance. Sentiment analysis enables these companies to monitor the various social media sites in real-time and act accordingly.

Sentiment Analysis analyzes and processes the emotions in textual data. In other words, the Sentiment Analysis determines if the provided information is favorable, negative, or neutral towards a certain topic or product. Sentiment Analysis employs techniques from NLP(Natural language processing), AI (Artificial Intelligence), and ML (Machine Learning) to process textual data. In machine learning, Sentiment analysis breaks down and characterizes human views, feelings, sentiments, and so on regarding a certain subject that is transmitted as either content or speech.

Sentiment analysis is a machine learning technique that examines the text for polarity, ranging from positive to negative. Machines learn how to identify sentiment without human intervention by training machine learning techniques with instances of emotions in text. Machine Learning for sentiment analysis starts with the collection of datasets. This dataset may be noisy, therefore it should be pre-processed using various Natural Language Processing (NLP) approaches. The Sentiment Analysis classification framework is illustrated as follows,

1. **Preprocessing**

Unanalyzed data is handled for feature extraction during preprocessing. It is further subdivided into the following steps:

Data cleaning: it is a process of finding and removing erroneous, ineffective, abstract content that does not depict any sentiment (like HTML tag or URL) and corrupt records from the Amazon book review dataset

Punctuation and special character removal: In writing, punctuation such as full stops (.), commas (,), and brackets () are deleted since they do not convey any feeling. Special characters such as "%", "#", and "$" used to signify cost, percentage, and comment are also removed.

Stemming: it is the process of reducing a token or a word for its root. For example, the root word or stem of terms such as "satisfaction," "satisfied," and "satisfying" is "satisfy.".

Stop word removal: Primarily prepositions, articles, and conjunctions such as the, is, at, which, and on that appear repeatedly in a phrase but do not convey any meaning. As a result, these stop words, which have no significance on their own, are deleted.

Case Normalisation: The words in the dataset are changed to lowercase or uppercase. We remove noisy data by imposing various preparation approaches. Furthermore, the preprocessed text is provided with an appropriate representation for categorization.

1. **Classification**

The Bag-of-Words approach is straightforward, it constructs a vocabulary from a corpus of texts and counts the number of times the words appear in each document. In other words, each word in the vocabulary becomes a feature, and a document is represented by a vector with the same length as the vocabulary (a "bag of words").

TF-IDF is an initials for term frequency-inverse document frequency, which provides a weight to a word based on the number of occurrences in the document as well as the frequency of the word in all documents. This reduces the weight of words that appear too frequently in all phrases, such as 'a,' 'the,' and 'as,' and enhances the weight of words that are significant in a sentence.

Word2vec is a collection of linked models used to generate so-called word embeddings. These are shallow, two-layer neural networks that have been taught to recover word linguistic contexts. Following training, word2vec models may be used to map each word to a vector of hundreds of elements that represents that word's relationship to other words.

Naive Bayes is a probabilistic classifier that employs the characteristics of the Bayes theorem while assuming substantial independence of the features. One benefit of this classifier is that it takes a little quantity of training data to generate the prediction parameters. Because of feature independence, just the variance of the feature is calculated rather than the entire covariance matrix.

Experimentation details

In this project, I used sentiment analysis of Amazon kindle store reviews. I’m using Kaggle for my dataset. First of all, I used the Amazon JSON dataset on <https://nijianmo.github.io/amazon/index.html#subsets> but the size is too big so I try to search anything about Amazon sentiment analysis review dataset on Kaggle. I chose this dataset because I think the dataset is not really complicated to understand rather than other datasets like music, movies, etc. I choose sentiment analysis topics to determine if reviews were positive or negative. It is difficult to read hundreds of evaluations on a regular basis and comprehend which products are loved or disliked by customers. It takes too much time and typically does not provide a concise overview of product comparison. I created a model that can predict if a review is favorable or bad based on the language, saving us time.

The dataset contains product reviews from the Amazon Kindle Store category from May 1996 - July 2014. There are a total of 982619 entries in this database. In this dataset, each reviewer has at least 5 reviews, and each product has at least 5 reviews.

First, I examine the dataset, understand the column in each category, and clean the data to evaluate data point distributions. I examined each column separately and compared common terms in positive and negative evaluations. I drop 22 null data in the row and the dataset becomes 982597 entries. Before preprocessing data, I check the overall column value and decide 1-4 stars labeled 0 which means negative sentiment, and the rest labeled 1.

The first challenge with this data is to remove irrelevant elements for modeling such as punctuation and upper-case characters using preprocessing functions. After preprocessing data, we use the cleaned data and prepare the data for modeling. Machine learning algorithms cannot operate directly with raw text. Rather, the words must be transformed into numerical vectors. Converting text to numbers is a critical step in predictive modeling. I use CountVectorizer and TfidfVectorizer in the Scikit-learn library to convert texts into fixed-length number vectors.

For the train and test model, I use the Naive Bayes technique to classify the dataset. From the beginning, my plan is to use 2 classifiers Naive Bayes and 1D CNN. However, I was unable to create the 1D-CNN model. The accuracy for 1D CNN is only approximately 30%, and I attempted to improve it while reducing the learning rate to 0.0001. So I decided to stop it and only focus on the Naive Bayes classifier. Using Naive Bayes classifier in count vectorizer I got 74% for testing accuracy and 82% for training accuracy. In TFIDFVectorizer the training accuracy decreases to 75% and 73% in testing accuracy

Discussion

Sentiment analysis is a crucial component for any issue (policymakers, stakeholders, corporations, etc.) to undertake many types of tasks such as predicting financial performance, understanding customer opinion, providing early warnings, defining election results, and so on. In the marketing sector, it is used to build strategies, analyze customers' sentiments about goods or brands, how people respond to campaigns or new launches, and why consumers do not buy certain items.

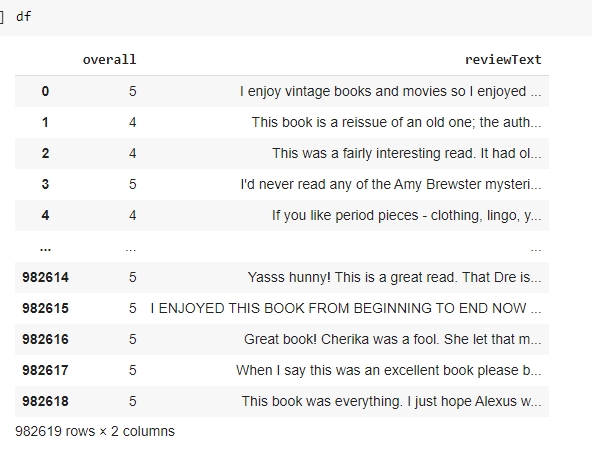
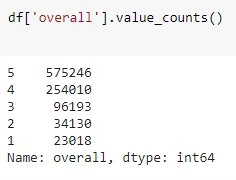
People typically look for items such as cell phones, laptops, and televisions, and before purchasing them, consumers may investigate the quality of the product that they are about to buy through reviews. In contrast, consumers cannot determine the quality of experience items such as movies, music albums, or books before using them, and therefore require more information than users of search products. People make purchasing decisions based on reviews that are regarded as the perceived value of a specific product.

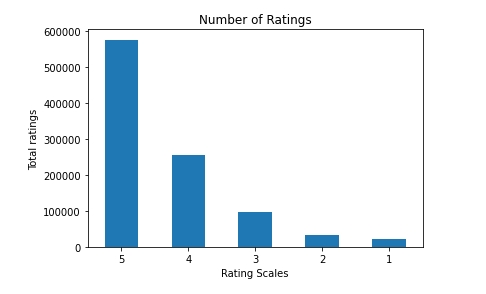
Using sentiment analysis in business may help us understand more about user requirements, adapt products and services, enhance customer support, and respond to social media in real-time and overtime. Sentiment analysis is one of the metrics that can help you improve your marketing plan because it affects many different areas of your brand's online visibility. Analyzing brand sentiment may help you understand the motives behind your consumers' purchase decisions as well as the intent behind their searches. Sentiment research can provide you with facts on which to base your marketing efforts. Sentiment analysis can provide you with data that will allow you to take your marketing approach to the next level.

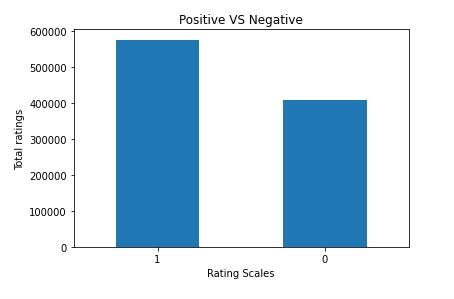
Sentiment analysis is still a newcomer to marketing analysis and still struggling with sarcasm and irony, although NLP, AI, and ML algorithms are improving in identifying distinct tones of Internet remarks. But still, the advantages of sentiment analysis are obvious. In the long term, sentiment analysis will improve your marketing performance.

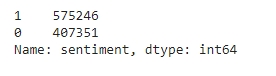
Attachment

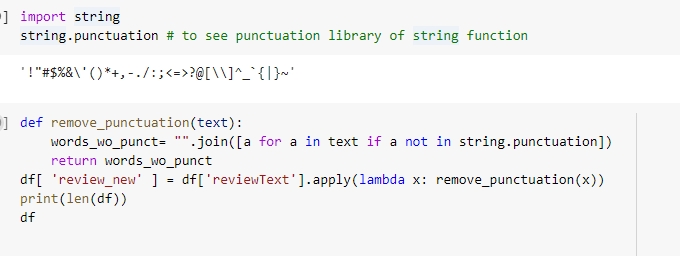


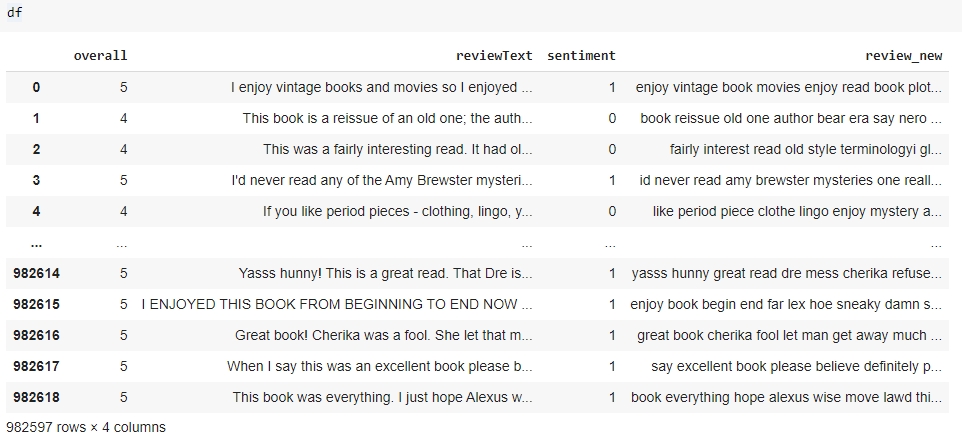
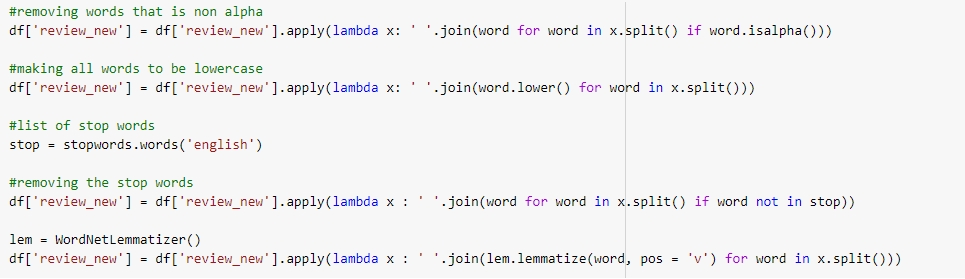


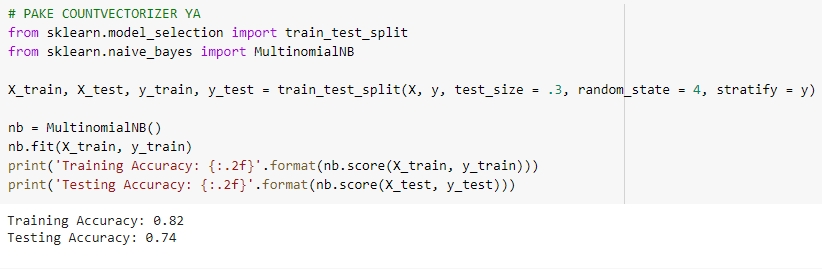












Resources

1. P. Balaji, O. Nagaraju and D. Haritha, "Levels of sentiment analysis and its challenges: A literature review," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 2017, pp. 436-439, doi: 10.1109/ICBDACI.2017.8070879.
2. Jagdale R.S., Shirsat V.S., Deshmukh S.N. (2019) Sentiment Analysis on Product Reviews Using Machine Learning Techniques. In: Mallick P., Balas V., Bhoi A., Zobaa A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore. <https://doi.org/10.1007/978-981-13-0617-4_61>
3. <https://monkeylearn.com/blog/sentiment-analysis-machine-learning/>
4. <https://www.linkedin.com/pulse/count-vectorizers-vs-tfidf-natural-language-processing-sheel-saket>
5. Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Scott D, ed. Proc. of the ACL 2004. Morristown: ACL, 2004, pp. 271−278. 
6. <https://brand24.com/blog/the-benefits-of-sentiment-analysis/>
7. Mohan Kamal Hassan et al 2017 IOP Conf. Ser.: Mater. Sci. Eng. 263 042090