Progress report for Sentiment analysis of hotel review

The paper title is **sentiment analysis of online review**.

The issue or the problem: the hotel industry is highly competitive in that hotel firms offer essentially homogeneous products and services, in the business field it's important to know about the previous experiments of the competitors in the field and to know about the user previous experiments so you can enhance your business to fit users expectations, by the huge amount of the data we need to use the big data techniques to implements operations on the data we obtained.

The data set I'm working on is **Hotel Reviews Data in Europe** this dataset was scraped from Booking.com, the biggest travel agency for reservations on any travel products you want including the hotel's rooms, the website has over 28 million reported accommodation listings, the data contains 515,000 customer reviews and 1493 luxury hotels across Europe, Meanwhile, the geographical location of hotels are also provided for further analysis.

My initial solution for the problem of the huge dataset is to split the dataset and process every part in parallel, this process could be implemented using big data techniques, I'm using the apache spark framework to control the clusters and the workers for this task, Apache spark scalable machine learning library (MLlib) will be used with different classification techniques from the library.

The related papers for this topic:

- **The first paper is Customer Experience and Satisfaction of Disneyland Hotel through Big Data Analysis of Online Customer Reviews:**

this study attempted to find the underlying dimensionality in online customer reviews reflecting customers experience in the Hong Kong Disneyland hotel and identified its relationship with customer satisfaction.

This paper applied three Models for data analysis

1. Semantic network analysis by Netdraw
2. factor analysis
3. linear regression analysis by SPSS 26.0

the goals of this research are to explore and demonstrate the utility of semantic network analysis, which is one of the important sections of big data analytics, to understand texts in the research of theme hotels, and then to identify to what extent it can help us to understand

customer experience through their online reviews with semantic network analysis.Google travel was used to collect relative online customer reviews, which is a trip planner service developed by Google for the web. The data collection process was conducted by SCTM3

The data collection period set in this research was four years, which is from 1 January 2017 to 4 May 2021, and “Disneyland Hotel” + “Hong Kong” were used as keywords for data collection since Hong Kong Disneyland hotel is a typical theme hotel owned by the Disney Company, 1493 reviews were collected with textual reviews, numerical ratings (from 1 to 5), reviewers’ information, and review date. Ucinet 6.0 packaged with Netdraw was adopted for data analysis and visualization for results of data analysis.

Freeman’s degree centrality and Eigenvector centrality were performed to measure how close a word is to the center in a network to conduct the analysis of semantic network of these top frequency words.

As a results, the centralities (Freeman’s degree and Eigenvector centrality) of 40 top frequency words were calculated and compared with the words’ frequencies

- **The second paper is A Large-Scale Sentiment Data Classification for Online Reviews Under Apache Spark**

This paper uses Apache Spark's scalable machine learning library (MLlib)

Also there is three classification techniques from the library are applied. These techniques are :

- Naive Bayes
- Support vector machine
- Logistic regression

The results are evaluated by performing the accuracy metric

This research aims to provide new experiments of sentiment classification on large-scale data using the Spark's MLlib by applying different MLlib classification algorithms and evaluating their performance.

The steps of the used approach is:

- data preprocessing
- Feature extraction
- Applying the machine learning classifier (Naïve Bayes, Support vector machine and logistic regression) under Spark environment
- Evaluate the result using the accuracy metrics

The data set used for experiments is the Amazon review polarity dataset, This dataset spans for 18 years. The dataset includes 35 million reviews , it also include class, review title, and review text columns. It contains 1,800,000 training samples and 200,000 testing samples in each class.

In the feature extraction phase the text is converted into a feature vector to be suitable for the classifier to deal with it, The Term Frequency-Inverse Document Frequency (TF/IDF) is

used.

- **The tird paper is Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews**

This research use the Long-Short Term Memory (LSTM) model and the Word2Vec model, the variables that used from the previous models in this research are:

- Word2Vec architecture
- Word2Vec vector dimension
- Word2Vec evaluation method
- Polling technique
- Dropout value
- Learning rate

On the basis of an experimental research performed through 2500 review texts as dataset the best performance was obtained that had accuracy of 85.96%.

The parameter combinations for Word2Vec are:

- Skip-gram as architecture
- Hierarchical Softmax as evaluation method
- 300 as vector dimension

Whereas the parameter combinations for LSTM are:

- dropout value is 0.2
- pooling type is average pooling
- learning rate is 0.001

Sentiment assessment : is used to gather and review viewpoints about products and services expressed in Tweets, reviews, comments, or blog posts.

Sentiment assessment research considers the hotel reviews for the purpose of the study. It classifies them into 2 categories.(positive or negative).

based on various factors:

- Services
- Prices
- Location
- Food
- Facilities

In the field of sentiment assessments different classical machine learning techniques has been applied:

- such as Naïve Bayes
- Support Vector Machines

- Logistic Regression
- Latent Dirichlet Allocation

Sentiment analysis research for reviews in Indonesian language has also been carried out using deep learning technique, i.e. the CNN (Convolutional Neural Network)
The problem of using the CNN:
CNN has the drawback that it cannot work with long sequential data.
It is because CNN does not possess a memory, so it cannot retain information about the word meaning.
So we using the LSTM (Long-Short Term Memory) model.
LSTM : is a kind of RNN (Recurrent Neural Network) architecture which is designed to "retain" values that have been obtained before for a specific period.
LSTM consists of 3 gates :

- input gate
- forget gate
- output gate

Word2Vec generates a vector space obtained from the corpus, which consists of words that are similar in the corpus and are adjacent to one another in the Word2Vec space.
the purpose of this research is to analyse sentiments present in the hotel reviews using LSTM model and word embedding by using the Word2Vec model, particularly for Indonesian language hotel reviews.
The methods of this paper divided into two phases:

1. Preparation of dataset
    a. data gathering
    b. data pre-processing
    c. Training using the Word2Vec
2. Creation of the sentiment division model
    a. Dataset partition
    b. Training and testing of LSTM

- The dataset used is a data of hotel reviews from the Traveloka website
- The dataset collected using Selenium and Scrapy libraries by automatic crawling
- The dataset collected from getting the cities then the hotels in this city then the comment section for every hotel
- This dataset contains 2500 review (1250 positive, 1250 negative)

The Word2Vec training process helps make the system learn vector representations of words by using the framework of neural networks

The input data is the pre-processed hotel review data.
whereas the output is in the form of a vector representation of every word.
The initial process during the formulation of a Word2Vec model comprises the building of vocabulary using input data.
in this study Word2Vec employs 12 different combinations.
The Hierarchical Softmax comprising the vocabulary set having a wordcount (W) is represented using a binary tree
a good model should be able to differentiate the real signal and the fake signal through logistic regression techniques. While the selection of word2vec dimension is based on the typical interval between 100-300.
lower than 100 dimension will lose properties of high dimensional spaces.
LSTM is a specific variant belonging to the Recurrent Neural Network (RNN) method.
The information to be updated and fed to the memory cell is determined by the gate input.
The forget gate is responsible for determining if the input/output information if suitable for passage.
The cell state is unaffected by the gate output.