SYNAPSE TASK 2: Natural Language Processing

Introduction

- Importance of rating system by using reviews

1. Customers opinions are very important for making any important decision.
2. It is necessary to understand your customers better and improve customer service.
3. This allows customers to have voice and create customer loyalty.
4. With this manner consumers are doing marketing for you.
5. People always go for summary rather than elaborate statements .In this way a rating system will be very helpful.Online food delivery comes with its own caveats. One of the biggest challenges is verifying the authenticity of a product. Is it as good as advertised on the application?

- Basis of rating system

1.The two techniques that we will use in finding opinions in text and detect emotions are sentiment analysis and opinion mining. This has become a hotspot in Natural Language Processing and retrieving information.Sentiment words and phrases are main indicator of sentiment classification.
2.Early work in Sentiment was on knowledge based or lexical based e.g., WordNet or Sentiwordnet. Classifying the sentiment of texts based on dictionary defines the polarity confidence of the words. Recently there have been some studies that take different machine learning approaches and build text classifies such as decision tree, Naïve Bayes and Support Vector Machine etc.
3.The proposition is based on sentiment analysis and an optimized probabilistic approach described by a group of researchers.

Title

Given a corpus of reviews for a restaurant on a food delivery app, analyze the sentiment behind each review. Classify the reviews into positive and negative and use it to give a rating (out of 5) to the restaurant.

Objectives

Conduct Sentiment analysis on the reviews of the users of food delivery app and numeric rating will be generated from polarity of content of reviews.
The proposition is based on sentiment analysis and an optimized probabilistic approach described by a group of researchers.

Study Material

1. Sentiment Analysis Based Product Rating Using Textual Reviews. International Conference on Electronics, Communication and Aerospace Technology ICECA 2017
2. Numeric Rating of Apps on Google Play Store by Sentiment Analysis on User Reviews. International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) 2014
3. An NLP Approach to Mining Online Reviews using Topic Modeling

Method

1. Gathering reviews from standard data set and analyze text

   - Gathering the information used techniques like pronoun resolution, entity extraction and con- referencing to segment each sentence. The reviews of the different materials such as cell phones ad laptops their corresponding rating is also given in the separate attribute.. By using the supervised learning techniques classifying them according the many pre-processing techniques. We will be using sentiment analysis to analyze the reviews.

- They extract sentiment expressions from diverse writing corpus and rank them with polarities. Much progress has been made in this field of sentiment analysis. However, most of the earlier works are based on summarizing a number of comments or reviews. Some of the important works are done using statistical techniques, based on machine learning or word-correlation algorithms.
- Before performing sentiment analysis the data is subjected to many pre-processing techniques and then identifying opinion data in the reviews and classifying them according to their polarity confidence i.e., whether they fall under positive or negative or neutral connotation. The open source data tool analysis tool called rapid miner is used to perform the step by step explanation of review processing.
- Sentimental analysis :Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.
- Apart from topic modeling, there are many other NLP methods as well which are used for analyzing and understanding online reviews. Some of them are listed below:

1. Text Summarization: Summarize the reviews into a paragraph or a few bullet points.
2. Entity Recognition: Extract entities from the reviews and identify which products are most popular (or unpopular) among the consumers.
3. Identify Emerging Trends: Based on the timestamp of the reviews, new and emerging topics or entities can be identified. It would enable us to figure out which products are becoming popular and which are losing their grip on the market.
4. Topic Modelling: Topic Modeling is a process to automatically identify topics present in a text object and to derive hidden patterns exhibited by a text corpus.

2. Preprocessing the data

- The raw data is preprocessed before extracting the subjective features includes removing HTML tags, removing missing attributes, replacing missing attributes with the another variable in order to avoid the fatal error. The data sheet contained comments expressed in English Language. For this the English responses are studied and extracted and each response was stored as a separate csv file. N gram Stemmer is a language independent Stemmer. It is a set of n in which following characters extracted from a word. “TEXT MINING” results in the generator of the diagrams. *T, EX, XT, TM, MI, IN, NI, IN, NG, G* are the tri grams. * denotes the padding space. There are n+1 such diagram and n+2 trigrams in a word containing n characters.

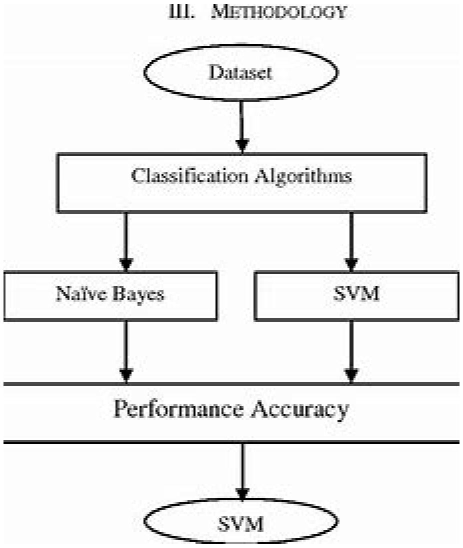
- The response statements were taken through pre-processing steps of tokenization, stop words removal and stemming Using appropriate operators in Rapid Miner. Word vector representations were also generated using the ‘TFIDF’ method. All the mentioned processes of data pre-processing and word vector generation were achieved using the ProcessDocumentFromFiles operator in Rapid Miner which can contain all the other required operators.

3. K-Means Clustering

- The clustering technique which divides the data into predetermined number of clusters. Whenever there are missing values in columns these algorithm replaces missing column with the categorical values with the mode and missing numerical values with the mean.
- $$J(V) = \overset{c}{\underset{i=1}{E}} \; \overset{ci}{\underset{j=1}{E}} \; (|Xi-Vj|)^2$$
- |Xi-Vj| is Euclidean distance between Xi and Vj. Ci is the number of data points in cluster. C is the number of cluster centre.

4. Opinion mining or sentiment analysis

- Opinion mining, or sentiment analysis, is a text analysis technique that uses computational linguistics and natural language processing to automatically identify and extract sentiment or opinion from within text (positive, negative, neutral, etc.).
- Polarity Detection
  1. Supervised methods are Machine learning approaches in which a classifier is trained based on a feature set, using labelled training data.
  2. Unsupervised Methods are mostly based on a Sentiment Lexicon in which each sentiment bearing word is associated with either a sentiment score or a set of sentiment bearing seed words as explained
  3. Hybrid Methods adopt a combination of both of the above mentioned categories to perform opinion mining.
- Support Vector Machine The SVM algorithm is mainly used to classify the text into positive and negative words. Firstly designing a hyper plane that classifies all training vectors into two classes.

- 

- G(x)>=1, Ax E class 1
- G(x)>=-1, Ax E class 1
- The total margin is computed by 1/\w\ +1/\w =2/\\w\\
- Minimize w is a nonlinear optimization task, solved by the karush Kuhn tucker conditions using LaGrange multipliers.
- Another method that can be used for segregation is Naïve Bayes algorithm.
  It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

4. Extracting Numeric Polarity

- Instead of extracting polarity directly a polarity probability for each candidate expression is determined first. To do this, for each candidate expression ci, two types of polarity probability will be calculated: P-Probability P rP (ci) that indicates the positive sensitivity and N-Probability P rN (ci) that indicates the negative sensitivity of the candidate expressions. These polarity probability will retain the characteristics such as P rP (ci) + P rN (ci)=1, that is the more like ci is positive or negative, the less likely it is negative or positive.
- Now we can define the consistent and inconsistent relations of two candidate expressions from the P-Probability and N-Probability of the candidate expressions. Let ci and cj are two candidate expressions whose polarity probability will be independent, their consistency probability will be P rP (ci)P rP (cj) + P rN (ci)P rN (cj) and the inconsistency probability will be P rP (ci)P rN (cj) + P rN (ci)P rP (cj). The consistent relation between ci and cj indicates that the review contains consistent sentiment, i. e. the expectation of consistent probability is 1.

- The difference between consistency probability and its expectation can be measured by squared error: (1−P rP (ci)P rP (cj)−P rN (ci)P rN (cj))2 in the network Ncons. Similarly, for the inconsistency network Nincons, the difference between inconsistency probability and its expectation is measured by (1 − P rP (ci)P rN (cj) − P rN (ci)P rP (cj))2. Therefore we get the sum of square errors (SSE) in both the networks.
- SSE = n −1 i=1 n j>i (wcons ij (1 − P rP (ci)P rP (cj) − P rN (ci)P rN (cj))2+ wincons ij (1 − P rP (ci)P rN (cj) − P rN (ci)P rP (cj))2)
- Where, wcons ij and wincons ij represents the weights of relation between ci and cj in the networks Ncons and Nincons respectively and n is the total number of candidate expressions. Here the SSE is used instead of absolute error so that the two types of relations can not cancel each other. It is expected that the P-Probabilities and N-Probabilities will minimize the SSE in such a way that the corresponding consistency and inconsistency probabilities will be closest to their expectations suggested by the networks. Now by replacing P rN (ci) with 1 − P rP (ci), P rN (cj) with 1 − P rP (cj), P rP (ci) with xi and P rP (cj) with xj .
- We get the following function:

minimize{ n −1 i=1 n j>i (wcons ij (xi + xj − 2xixj) 2+ wincons ij (1 − xi − xj +
2xixj)
2)} where, 0 ≤ xi, xj ≤ 1 for i, j = 1, 2, ..., n

- Here, the candidate expressions lying in the root set will be assigned with 1 or 0 according to their predefined polarity. The probabilities of other candidate expressions can be found out by the above stated function. For this, the L-BFGS-B7 algorithm will be used to solve the above problem and the polarities will be generated. Finally, the candidate expressions will be assigned with a P-Probability and N-Probability.
- This P-Probability and N-Probability will determine the positive or negative sentiment of the user reviews. As the polarity gained by this will be in the range between 0 and 1, it will be converted to a rating on a scale between 0 to 5 as the prevailing starred rating on the food delivery app.

Conclusion:

- The sentiment expressions in user reviews varies diversely. This diverseness can be in single words like as awesome, worthy, etc., as well as in case of multi-word phrases like time wasting, great to use etc. Moreover, there can be mixture of formal and slang expressions too with abbreviations and spelling fluctuations. On a research done on 3000 tweets, it is seen that 15.25% expressions contains slangs [3]. The process have to be developed such that it can differentiate and handle with this diverse domain of sentiment expressions. Extracting sentiment expressions was done earlier from formal corpus. Generally the user reviews are not written following the formal rules. So, the exact sentiment expression cant be extracted from the user reviews with some predefined patterns. In a word, the writing style of the user reviews cant be easily parsed with various parsers and parts-of-speech taggers. These tools are typically made on the basis of standard spelling and grammar. Here, the solution to the problem is proposed using sentiment analysis on the user reviews and considering the starred rating. Sentiment analysis will be conducted on the reviews of the users and a numeric rating will be generated from the polarity of the content of the reviews.

- SVM's are very good when we have no idea on the data.
- Works well with even unstructured and semi structured data like text, Images and trees.
- The kernel trick is real strength of SVM. ...
- Unlike in neural networks, SVM is not solved for local optima.
- It scales relatively well to high dimensional data
- SVM and Naive Bayes algorithm are precise and more accurate for opinion mining. Therefore they are widely used to generate rating based on reviews.