Sentiment analysis using naïve bayes classification and svm: a case of skytrax da...

Business, Strategy

\n[toc title="Table of Contents"]\n

\n \t

\n[/toc]\n \n

Data mining tasks such as classification or prediction are applied in various domains including manufacturing and business. Sentiment Analysis has significantly fledged over recent years in our day-to-day engagement information and technology. As a field of study, sentiment analysis research has association with aspects of artificial intelligence research including Business Intelligence.

This project is to demonstrate how Sentiment Analysis can be applied to Business Intelligence. Sentiment analysis, also known as opinion mining, serves to determine the inclination or attitude of a communicator through the contextual polarity of their writing or speaking. The term " Sentiment Analysis" tells that it is analysis of the various sentiments expressed by humans on a specific topic/subject of discussion, or the opinions of/feedback given by customers to various business organizations. Sentiment analysis is the practice by which information is extracted from the opinions, appraisals and emotions of people in regards to entities, events and their attributes. This field is also characterised as a sub-category of text-mining, which falls under Natural Language Processing and Data Mining.

Text Mining as defined by (Al-Azmi, 2013) deals with textual data rather than records; and it differs from data mining in terms of methodology and techniques used, such as Text Mining using complex Natural Language Processing (NLP) techniques. In the 1980's, computer processing power in conjunction with machine learning capabilities led to a breakthrough in NLP which enabled the exponential growth in the abilities of machines to ' appear' intelligent (Badenhorst & Fitzgerald, 2012). Business Intelligence according to (Costa & Souza, 2012) is made up of two steps. First, collecting, transforming, and data loading from unstructured (e. g. social networks websites, emails) or structured (e. g. ERP, CRM) sources resulting in data warehouse which includes repository of data that is integrated, topic-oriented, time variant, and nonvolatile. Second; using analytical tools for the propagation and analysis of knowledge (Knowledge Discovery Process).

Businesses need to have a complete understanding of their customer's opinions and needs on their products or services, but they face the problem of dealing with unstructured text from source of customer's opinions and need. Consumer products and services sentiments have now become more than just a source of customer reviews and references but a source for customer services, business intelligence, and product brand reputation management. Voice of the customer on product/service reviews and customer feedback bring forward fundamental questions that need businesses to prioritise their efforts and resources in attending to. Sentiment analysis comes in as an effective tool in driving towards the Knowledge discovery in turn lead the organisation to Business Intelligence. Some of the questions in need of deriving sentiment are:

- Are the customers satisfied with services, products, and support?
- What customers think of products and services offered by competitors
- What influences the market and how opinions propagate.
- What do the customers like.
- What problems do customers have?
- And, what additional features would the customer like to have and are willingto pay for

The essence of this project is in the confines of demonstrating how supervised learning techniques such as Support vector machines and Naïve Bayes can be used to create sentiment analysis models which classifies the airline reviews as positive, negative opinion or neutral. The data set of reviews is collected from Skytrax, a London based organization which specializes in Research and Quality Advise provision to the air transport industry, advising airlines and airports around the world for quality improvement and quality leadership issues (Skytrax, 2018).

Background and related workHumans have the innate ability to determine sentiment however this process is time consuming, inconsistent and costly in a business context especially when the sentiments of goods/services have to be derived from a huge collection set of reviews and feedback on social media platforms, emails, blogs, etc. It is not realistic to have people individually read tens of thousands of user or customer reviews and then score for sentiments therefore sentiment analysis using machine learning techniques becomes relevant in that part of the business as has been applied to areas such as Twitter sentiment analysis, movie review sentiment analysis. Twitter has arisen as one of the most typical example of social media platforms where scholars, advertisers and political activists put forth their views and thoughts in the form of status messages which are also termed as “ tweets”. There are occurrences where real time Twitter sentiment analysis was performed; an example was 2012 U. S. Presidential Election. (Wang, et al. , 2012) collected public reviews from Twitter about the U. S. presidential candidates for 2012 elections and used as inputs which were passed through a sentiment model which was designed to determine tweets sentiments (positive, negative, neutral, sarcastic, humorous or unsure).

The motivation behind this study was to comprehend how these electoral events have an effect on public opinions. Sentiment Analysis has also been used to classify movie review remarks obtained from social networking site ’Diggs’ into positive/negative/neutral and subjective/ objective with the help of machine learning techniques (Yessenov & Misailovi, 2009).

Statement of the problem

The challenge to be addressed in this project is the limitation to accurately classify whether an expressed opinion in a document, a sentence, or an entity feature/ aspect is positive, negative or neutral.

Aim

The main goal of this problem is to develop a sentiment mining model that can process and score sentiments of a given text.

Objectives

1. To implement a machine learning algorithm to perform sentiment analysis.
2. To implement three classes of sentiments namely positive, neutral and negative.
3. To achieve 80% or more in classification accuracy.
4. To build a graphical user interface for the visualisation purposes of the classified sentiments. JustificationBusiness Intelligence as a sub type of sentiment analysis offers arrange of benefits and advantages that is just focused to competitive advantage, but a more varied set of advantages that are worthy of consideration to any company. Some of the advantages as stated by (Al-Azmi, 2013) below:

Competitive Advantage:

- market research; finding elements of market dominance
- risk management; bankruptcy prediction, better investments
- manufacturing optimization; better material usage, shipments, scheduling, etc

Sentiment Classification Techniques (Symeonidis, 2018)Machine Learning is usually used to classify sentiment from text. This technique involves statistical models such as Support Vector Machine (SVM), Bag of Words and

Naïve Bayes (NB). SVM and Naïve Bayes methods are going to be used to develop sentiment mining models. SVM is primarily a method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVMs perform both regression and classification tasks and can handle multiple continuous and categorical variables. Naïve Bayes classifiers mostly used in text classification (due to better outcome in multi class problems and independence rule) have higher success rate as compared to other algorithms (Ray, 2011). As a classification problem, sentiment analysis uses the evaluation metrics of Precision, Recall, F-score, and Accuracy. Also, average measures like macro, micro, and weighted F1-scores are useful for multi-class problems. Depending on the balance of classes of the dataset the most appropriate metric should be used.

The most popular techniques include:

- Remove numbers
- Stemming
- Part of speech tagging
- Remove punctuation
- Lowercase
- Remove stop-words

The Skytrax data set of Airline reviews by customers is going to be used for training and testing the models. The programming framework is mainly going to be Python. The SDLC procedure to be followed is the Agile iterative methodology. Agile methodology is a practice that promotes continuous iteration of development and testing throughout the software development lifecycle of the project (Guru99, 2018). Both development and testing activities are concurrent unlike the Waterfall model. In iterative development, feature code is designed, developed and tested in repetitive cycles. With each reiteration, additional features can be designed, developed and tested until there is a fully functional software application ready to be deployed to customers. Agile iterative methodology is preferred because of the need of iteration in training and testing the SVM algorithm until it achieves at least 80% accuracy in the classification of sentiments.

Scope

The essence of this project is to demonstrate how sentiment analysis can be used as an effective tool for scoring sentiments on reviews, feedbacks, opinions and drive to knowledge discovery in the business processes. The constraints of this project are set on the definition of Business intelligence as a phenomenon of propagation and analysis of knowledge. This involves the use of appropriate tools to build, train and test a model for sentiment analysis and lastly the visualization of the analyzed data. Expected resultsA machine learning model for sentiment analysis that can classify a comment, review, phrase of words into either positive, negative or neutral sentiments with positive score meaning “ brilliant effort, loved your work”; negative score meaning “ totally dissatisfied with the service” and the neutral score meaning “ good job but I will expect more in the future”. Also, a graphical user interface that will accept user input of a phrase to be analyzed and scored for sentiment.

Lastly the project should be able to put the analyzed data for visualization to aid in the utilization of the knowledge that was created. Proposed timeline This project is going to follow the SDLC phases and therefore all the constraints of each stage in the life cycle are going to determine the time taken to achieve the aim and bring the project deliverables. Although, speculatively the project will not take not more than 8 months.