

REVIEW-1

ITE2006 DATA MINING TECHNIQUES

RegNo:- 19BIT0096

Title:- Amazon Review Classification

Name:- Byreddy Joseph Prasanth Kumar Reddy

Slot:- D2+TD2

Abstract:

The world we see nowadays is becoming more digitalized. In this digitalized world e-commerce is taking the ascendancy by making products available within the reach of customers where the customer doesn't have to go out of their house. As now a day's people are relying on online products so the importance of a review is going higher. For selecting a product, a customer needs to go through thousands of reviews to understand a product. But in this prospering day of machine learning, going through thousands of reviews would be much easier if a model is used to polarize those reviews and learn from it. We used supervised learning method on a large-scale amazon dataset to polarize it and get satisfactory accuracy. Reviews on Amazon are not only related to the product but also the service given to the customers. If users get clear bifurcation about product reviews and service reviews it will be easier for them to take the decision, in this paper we propose a system that performs the classification of customer reviews followed by finding sentiment of the reviews. A rule based extraction of product feature sentiment is also done. Also we provide a visualization for our result summarization.

Introduction:

Amazon is one of the largest online vendor in the World. People often gaze over the products and reviews of the product before buying the product on amazon itself. But the reviews on amazon are not necessarily of products but a mixture of product of product review and service review. The buyer is misled as the overall sentiment (rating classification) that amazon gives is a collective one and there is no bifurcation between a service review and product review. The proposed model satisfactorily segregates service and product review, in addition to this it also classifies the review as Feature review if the user talks about some particular product feature. A featured review is nothing but a product review, our model also gives sentiment of the text about the product feature. For example, if the user writes in his review, "the camera for this phone is very good.", then we also classify camera feature as positive. We aim to build a system that visualizes the review's sentiment in the form of charts.

Literature Survey:-

1. Research Based on Different range of Algorithms [2016]:

In our work we have used reviews of iPhone 5 extracted from Amazon website. We studied all the reviews and got to know that there are many reviews in which the user talks about the service provided by amazon and its sellers. So we decided to classify reviews into service, product and feature based reviews. We also found that the sentiment of each review is very obvious, the review rating provided by the user mirrors what the user writes as his/her review, i.e., if the user writes something bad definitely the overall rating the user gives is either 1 or 2 out of 5. This is from our study of a set of amazon reviews on iPhone 5. Many algorithms can be used in opinion mining such as Naive Bayes Classification, Probabilistic Machine Learning approach to classify the reviews as positive or negative, have been used to get the sentiment of opinions of different domains. Our work also concentrates on feature extraction and finding out the sentiment of the particular feature. We have used POS tagging technique on sentence level. In our approach we have made certain rules using the tags of particular word and using list of words with respective sentiment value to find the feature and then getting the appropriate sentiment from it. The Sentiment model that we have proposed is designed based on the uncertainty of the amazon reviews. Our work also include summarization in the form of charts for overall view of the sentiments of the users on the product or a particular feature.

2. Based On POS Tagging Technique [2015]:

Our work mainly concentrates on feature extraction and finding out the sentiment of the particular feature. We have used POS tagging technique on sentence level. In our approach we have made certain rules using the tags of particular word and using list of words with respective sentiment value to find the feature and then getting the appropriate sentiment from it. The Sentiment model that we have proposed is designed based on the uncertainty of the amazon reviews. Our work also include summarization in the form of charts for overall view of the sentiments of the users on the product or a particular feature. The portion of expression of each phrase in a sentence known as POS (component-of-speech) marking. It is one specific feature of natural language processing. For the extraction of knowledge, POS tagging is so critical since each group has a particular function in a paragraph. Nouns assign titles to our world's things or concepts. Some adverbs may be precisely the same as an adjective. Apache Open NLP tagger is used in the proposed method to tag the extracted analysis word.

3. Based on Sentimental Analysis [2017]:

Sentiment analysis and opinion mining work started in 1966 as the General Inquirer program was developed. Sentiment analysis is a natural language processing problem, also known as opinion mining (NLPs), which implies the detection and retrieval of contextual text source knowledge. Sentiment analysis indicates the emotions conveyed by a consumer on a piece of text, and it is used to interpret user reviews and comments from social media contents. The purpose of sentiment classification is to interpret and identify written comments of users into positive or negative attitudes such that the program may not need to grasp the meaning of sentence or text entirely. The authors stated that sentiment analysis has a variety of applications to establish customer loyalty and satisfaction from their critical reviews and user comments. The analysis usually helps other technologies, such as recommendation systems to develop the framework. This project also studies product and time relationship evaluations along with syntactic and semantic relations. This will vastly boost the precision of the description of feelings. The point is that a common product may receive higher ratings and that a less popular product may receive higher ratings.

4. Research Based on Sentiment Polarity Methodology [2015]:

The service and the product review's polarity is the rating the user provides for that review. The Good Reviews are those with rating 5 stars and 4 stars, Average Reviews are those with rating 3 stars and Bad Reviews are those with rating 2 stars and 1 star. Finally, when a feature sentiment is extracted the sentiment phrase is sent to a polarizer method, this method basically returns +1 if the phrase is a positive sentiment else -1 if the phrase is a negative sentiment. Firstly, the phrases are tested for indirect opinions such as "Battery no better than iPhone 4s", the test phrase is tested for certain pre-defined phrases that were found during manual analysis of reviews. Next if the phrase test fails, the review is tested for the word "not" if the word not exists then everything after not is polarized meaning every word after not is tested for whether it is a positive word or a negative word and consecutive words polarity are added and finally negated, for example "Camera is not good" this phrase is classified as negative as the word "good" is negated by the word "not". Lastly if "phrase" and "not" test fail the test phrase is broken down into words and polarity of each word is found from a dictionary of sentiment words bifurcated as good and bad words and collective polarity is considered i.e., if the sum is below 0 the outcome is negative (-1) else outcome is positive (+1).

5. Research Based on Navies Bayes Classifier [2017]:

It is a method to approach the following segments of the review which the customer are providing such that the review are consisting of several un wanted words such that classifier are being in search of the tag words such that the defined tag words in the vectors which are positive and negative such that the there are several condition to calculate the following probability of calculating the points to determines about the classification of the major review where there are several special symbol which must be removed by the classifier. Thus the navies' Bayes classifier is an method and an approach of statistics to get through the vectors contain the word and several library package to determine that the particular tag word must be counted in order to give there result about an accuracy of data to which we can determine whether the user can be able to identify the review rating about user reviews. Naive Bayes is used as the classifier. They over that naive Bayes offers higher result for word stage than divorced word and multiword. The Principle cons of this paper are, they used the best naive Bayes classifier set of rules from that we are able to not get a comfortable result. In paper it's used less arduous algorithms therefore it's straightforward to know. The system offers exaggerated accuracy on SVM so it cannot design nicely on a huge dataset. They used assist vector tool (SVM), provision regression, selection trees method.

S.no	Research	Algorithm	Usage
1.	Research on Amazon I Phone 5 Reviews	Many Algorithms used including k-means, Naïve Bayes algorithm.	This can be used in opinion mining. Whether positive or negative.
2.	Research on Amazon Reviews based on Design.	POS Tagging Technique.	This one mainly concentrates on feature extraction and finding out the sentiment of the particular feature.
3.	Research on sentimental Analysis.	Natural language processing(NLP).	This implies the detection and retrieval of contextual text source knowledge.
4.	Research on Amazon Different types of Reviews.	Sentimental Polarity Methodology.	When a feature sentiment is extracted the sentiment phrase is sent to a polarizer method, this method basically returns +1 if the phrase is a positive sentiment else -1
5.	Research on most used Amazon products reviews	Navies Bayes Classifier.	Which the customer are providing such that the review are consisting of several un wanted words such that classifier are being in search of the tag words

Problem Statement:

The crux of this Amazon Review Classification problem is implementing the Nearest Neighbour algorithm but it presents the opportunity to choose from several information retrieval methods for text mining. The solution I implemented was a TF-IDF based model that vectorizes each of the 18506 Amazon reviews in train and test datasets and then uses Nearest Neighbour to classify them as positive (+1) or negative (-1).

Existing Methods:

Naive Bayes Algorithm:

Naive Bayes is an excellent Machine Learning algorithm when it comes to text classification. Classifying text data generally includes a lot of pre-processing in the form of text mining, before it is suitable for building models. Text mining is the process of transforming raw, unstructured text data into a meaningful, structured form in order to extract valuable insights. Here, we shall construct a simple Naive Bayes model which will take text reviews of Amazon products and classify them as either “Positive” or “Negative”.

K-means Algorithm:

K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups. You define the attributes that you want the algorithm to use to determine similarity. The k-means algorithm expects tabular data, where rows represent the observations that you want to cluster, and the columns represent attributes of the observations. The n attributes in each row represent a point in n -dimensional space. The Euclidean distance between these points represents the similarity of the corresponding observations. The algorithm groups observations with similar attribute values (the points corresponding to these observations are closer together).

Proposed Methods:

Natural Language Processing:

We used NLTK for preprocessing the train and test datasets. At the preprocessing stage, we chose to use lemmatization over stemming even though both these forms of term normalization yield similar results and depend more on the identification of “root” words rather than linguistic usage in a particular text document.

Term Frequency-Inverse Document Frequency (TF-IDF):

Sentiment analysis in this problem is reduced to knowing which words occur most frequently in positive reviews and which ones occur most frequently in negative reviews. TF-IDF involves two steps – one is computation of a normalized Term Frequency that is indicative of how often a word appears in a document. Since text is often unstructured, the term frequency of a word may change depending on the length of the document it appears in. For example, a word may appear more frequently in a longer document than a shorter one.

$TF(t) = (\text{number of times term } t \text{ appears in a document}) / (\text{total number of terms in the document})$

Inverse document frequency measures how important a term is. To avoid one pitfall of IDF calculations which is measuring the influence of stop words like “is”, “this”, “a”, “the”, “and”, etc., I have filtered out stop words from the corpus in the natural language processing stage. I intended to reduce the number of dimensions and calculations involved by reducing the vocabulary in this way. IDF is calculated as follows:

$IDF(t) = \log_{10}(\text{total number of documents}) / (\text{number of documents with term } t \text{ in them})$

TF-IDF vectorization of train data would give me 18506 vectors containing TF-IDF weights for each unique word in the training corpus.

	0	1	2	3	4	5	n
--	----------	----------	----------	----------	----------	----------	--------------	----------

0	0.456	0.0012	0.005	0.3	0.0001	0.1	0.232
:	:	:	:	:	:	:	:	:
18506	0.22	0.007	0.065	0.8	0.336	0.4785	0.009

Fig.1 Representation of a TF-IDF model for
train data

The immediately apparent drawback of this method is that the dimensions of the vectors increase with the size of the vocabulary. With 18506 reviews in the train file and an equal number in the test file, the size of the vocabulary including unique terms from both in our case was 36956.

To make similarity calculations between matrices of such sparsity feasible, I used Latent Semantic Analysis to reduce each review to two components.

Latent Semantic Analysis (LSA):

LSA converts term document matrices to semantic spaces of lower dimensionality. LSA is a technique used to combat synonymy and polysemy in document classification, clustering and information retrieval. Synonymy refers to the phenomenon where different words are used to describe the same idea. Polysemy refers to the phenomenon of a word having multiple meanings across subject and usage contexts. The assumption in LSA or similar dimensionality techniques is that the representation obtained is the “true” representation and it uncovers the underlying similarity between two documents which have no common terms. A disadvantage of LSA is that a significant amount of information is lost when term documents are reduced to a small number of components. Singular Value Decomposition (SVD) is used to extract features from documents in my solution.

Nearest Neighbour Algorithm:






The variant of the Nearest Neighbour algorithm I've implemented is called k-Nearest Neighbour (kNN). The kNN assigns a class label to an unseen data instance based on the class labels of the nearest 'k' seen instances. This is called majority voting. The similarity metric used is cosine similarity. Cosine similarity is calculated using L2 normalization as a parameter for SVD while creating the LSA pipeline in my solution. The computational complexity is important while setting a while for 'k'. A general rule of thumb seems to be that we can start with $k = n^{1/2}$ for n samples, which would be $k = 136$ for our 18506 samples. I've dealt with the problem of breaking ties during majority voting by selecting odd values of k while testing for the optimal k. In my last submission, $k=379$.

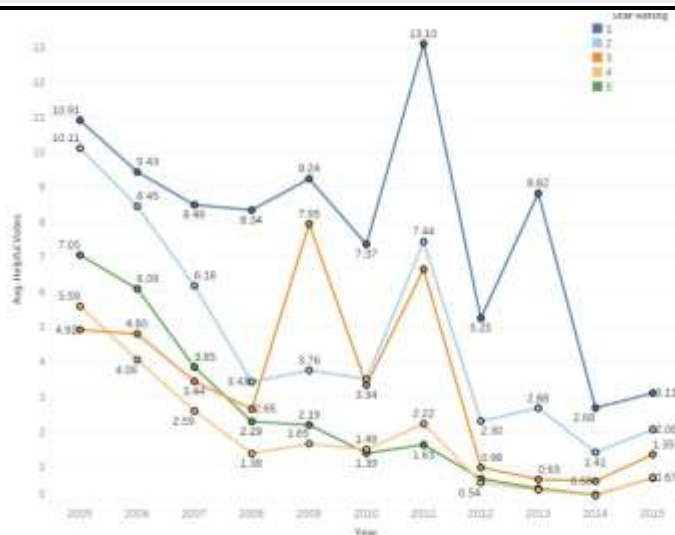
Datasets Description:

This dataset train.csv and test.csv contains product reviews and metadata from Amazon, including 18506 reviews as of 09/19/2016. This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features.

Sample Data:

Examples of some Sample data are,

Result Grid			Filter Rows:		Export:		Wrap Cell Content:		Fetch rows:
customer_id	product_id	review_body	helpful	overall	review_date	review_headline	star_rating		
10005833	B002A48...	Well deserved to be a ...	0	0	2015-08-2...	Five Stars	5		
10008659	0671733354	Perfect.	0	0	2015-08-2...	Five Stars	5		
10010780	1502525496	This book kept my inter...	1	1	2015-08-3...	Love it love it lo...	5		
10011040	0881030368	Enlightening....eye op...	0	0	2015-08-3...	Must read for e...	4		
10012167	1508973458	JR Harding has been a ...	0	0	2015-08-3...	WOW	5		
10014050	1579653510	we love to cook, but t...	2	3	2015-08-3...	we love to cook	3		
10014149	151482096X	Great book. The storie...	0	0	2015-08-2...	Love in Mistleto...	5		
10014701	0692406735	Received a copy of her...	0	0	2015-08-2...	In my reading s...	5		
10015224	0061258474	Stupid Wars is a non-fi...	0	0	2015-08-2...	Impressive Tak...	4		
10015224	0758203993	I purchased this book ...	0	0	2015-08-2...	Extremely Hard...	1		
10015224	1564144844	As an avid history-buff...	0	0	2015-08-2...	Some of my fav...	4		
10016045	0800721985	I received a copy of thi...	0	0	2015-08-2...	Choppy action	3		
10016045	085721604X	I received a copy of thi...	0	0	2015-08-2...	Very thorough, ...	4		
10016708	1608193942	In the same way that ...	3	3	2015-08-2...	All it takes is a l...	5		
10017695	1477816208	*I received a free cop...	1	1	2015-08-3...	*I really enjoy...	4		
10017822	0987650408	I've been using this bo...	0	0	2015-08-3...	Cautiously opto...	4		
1001811	0399536213	It was everything that ...	0	0	2015-08-2...	Books	5		
10018115	0887431488	Great for review.	0	0	2015-08-3...	Five Stars	5		
10018115	0938256343	Great for review.	1	1	2015-08-3...	Five Stars	5		
10018115	0938256467	Great for review.	0	0	2015-08-3...	Five Stars	5		
10018207	0991858891	Great book.....gr...	0	0	2015-08-3...	Five Stars	5		
10018207	1493010042	Great book.....great s...	1	1	2015-08-3...	Five Stars	5		
10018887	0692289771	Good format... easy to...	2	2	2015-08-3...	Great guide an...	5		
10020112	1514273934	Disappointed. Did not l...	0	1	2015-08-2...	Disappointed. D...	1		
10020322	1451666179	Best book ever.	0	0	2015-08-2...	Five Stars	5		
10020322	1451666179	Best book ever.	0	0	2015-08-2...	Five Stars	5		



References:

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. "Scoring, Term Weighting and the Vector Space Model." Introduction to Information Retrieval. New York: Cambridge UP, 2008. N. pag. Web.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. "Matrix Decompositions and Latent Semantic Indexing." Introduction to Information Retrieval. New York: Cambridge UP, 2008. N. pag. Web.

Alex9311. "Methods of Calculating Cosine Similarity between TF-IDF Vectors." StackOverflow. N.p., 2 Feb. 2016. Web. 19 Sept. 2016.

Null-Hypothesis. "Tf-idf-cosine: To Find Document Similarity." StackOverflow. N.p., 25 Aug. 2012. Web. 19 Sept. 2016.

Duran, Fletcher. "Dealing with Ties, Weights and Voting in KNN." StackOverflow. N.p., Dec. 2010. Web. 19 Sept. 2016.

S. ChandraKala¹ and C. Sindhu², "OPINION MINING AND SENTIMENT CLASSIFICATION: A SURVEY," Vol. 3(1), Oct 2012, 420-427.

G.Angulakshmi , Dr.R.ManickaChezian , "An Analysis on Opinion Mining: Techniques and Tools". Vol 3(7), 2014 www.iarcce.com.

Callen Rain, "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning" Swarthmore College, Department of Computer Science.

