Reviews Analysis of Online Retail Stores in UAE: Analytical Study of Sentiments Through Social Media

Riktesh Srivastava, Mohd. Abu Faiz

Abstract: Text mining for social media has now become a decisive tool for marketing, and many businesses understand the supremacy of embracing technology into their marketing campaigns. These texts are the “Consumer language”, owing to its spread and reach. There is no reservation that use of user generated texts have stimulated the companies to identify them and use it for decision making, however, classifying sentiment analysis through these texts is still a fresh sensation. Online retail companies in UAE are an early adopter of social media, but how do they use text mining techniques is still a matter to worry upon. The study proposes a model to collect reviews from multiple sources and identify sentiments and topics simultaneously. The model is the tested on 3 online retail companies in UAE and the results depicts productive outcomes.

Index Terms: Sentiment Analysis, Liu Hu algorithm, Plutchik modeling, Latent Semantic Indexing.

I. INTRODUCTION

Companies are eager to join the dialogue with consumers through social media [1], but are finding it difficult to increase the engagement. There is also a claim that social media acts as a tool for continuous interaction between consumers and companies [2], but actual interaction through social media is just 2% [3]. Many applications were then introduced to increase interactions:

- Twitter hashtags to get higher consumer engagement [4], [5]
- Facebook Ratings and Reviews, to post recommendations directly onto business pages [6]
- Blogs for broader consumer feedbacks [7], [8], and
- Review sites as an intermediary between consumers and companies, with 1/3 of consumers use these reviews before buying [9]

Consumers are using these applications for voicing their opinions, and thus analyses of these data has increased. The consumer can know the qualities of the product from the experiences shared by people on these applications, which can be useful for them before buying online. Online retail companies can improve their product or services on the basis of consumers reviews. The analysis of online contents to extract reviews requires deep understanding of natural text; abilities of most of the existing models are known to be inadequate.

The proposed model works in two fold, wherein, the model does both sentiment analysis and topic modeling of the extracted text. For sentiment analysis, the texts are first classified into positive and negative sentiments via Liu Hu Algorithm [10]–[12] and then divided into 8 polarities – Anger, Disgust, Fear, Sadness, Joy, Surprise, Anticipation and Trust respectively using Plutchik modeling.

Secondly, the model also categorizes 5 topics for the texts using Latent Semantic Indexing (LSI) algorithm. LSI is Natural Language Processing (NLP) algorithm that consents set of observations to be defined by unobserved groups [13], [14]. LSI is considered to be the most appropriate algorithm for identifying the topic representation of each texts [15]. Consequently, the proposed model generates three outcomes,

- Outcome 1: Sentiment Analysis (positive and negative sentiments) [as mentioned in Part 2a section]
- Outcome 2: Classification of Sentiments into Polarities [as mentioned in Part 2b section]
- Outcome 3: Identification of imperative topics [as mentioned in Part 2c section]

The paper is divided as follows: Section 2 explains the proposed model, with step-wise justification of each steps. Section 3 tests the model on extracted texts for 3 online retail stores in UAE. Section 4 does the analysis of outcomes acquired and concludes the paper.

II. PROPOSED MODEL

Figure 1 elaborates the proposed model for text analysis of the extracted text. As specified sentiment analysis and topic selection for the consumer feedback are collected from various sources regarding the 3 online retailers in UAE.

![Figure 1: Proposed Text Analysis Framework](image)

The collected texts are first preprocessed for analysis. The preprocessed texts serve as feeder to both sentiment analysis and topic modeling.
A. Preprocessing Steps

Preprocessing is a crucial step, since selecting the appropriate preprocessing methods, polarity can be clearly classified [16]–[18]. Preprocessing is used to eliminate the texts, as some texts include mishmash of English and other language words.

The task required us to classify a text into positive, negative and neutral polarity categories. This can essentially be treated as a 2-step process [19]

- Classify each text into subjective (positive/negative) and objective (neutral) classes.
- Classify subjective text into positive and negative ones

In order to follow the 2-steps, there will be 4 steps in preprocessing of texts(Srivastava, 2018), as the text may also include symbols, emoji, hyperlinks or other language texts, which needs to be removed and only intended text to be selected.

Step 1: Transformation

The text collected are in raw format (combination of short form of words, notations and other symbols). Transformation converts this raw format into more meaningful format. It includes conversion of texts in lowercase, remove accents, parse html and remove URLs.

Mathematically, transformation steps can be denoted by equation (1) as:

\[ \sum_{t=1}^{n} \prod_{i=1}^{t} i_{t_{i}} = \sum_{t=1}^{n} \prod_{i=1}^{t} i_{t_{i}} = \sum_{t=1}^{n} \prod_{i=1}^{t} i_{t_{i}} \]

where, \( t \) is number of texts which contain \( j \) words. Notice from equation (1) that each words of texts are not disjointed and are embodied as sentences.

Step 2: Tokenization

Tokenization is a process of breaking texts into words (as stated in equation (1)), called tokens.

Equation (1) can now be redefined as in equation (2) as:

\[ \sum_{t=1}^{n} \prod_{i=1}^{t} i_{t_{i}} = \sum_{t=1}^{n} \prod_{i=1}^{t} i_{t_{i}} = \sum_{t=1}^{n} \prod_{i=1}^{t} i_{t_{i}} \]

Step 3: Normalization

Step 3 receipts all words as cited in equation (2) as input and executes stemming and lemmatization. Stemming is heuristic progression of axing of derivational affixes. There are likelihoods that after stemming, some words may appear irrelevant, though relevant, and are likelihoods that after stemming, some words may appear irrelevant, though relevant.

Step 4: ngram

Next step is termed as ngram, which is sequence of \( n \) words. The combination of words is principally a set of co-occurring word, and, while computing the ngram typically move one word forward. The generic representation of ngram is:

\[ \text{ngram}_{k} = k - (t - 1) \]

where,

\[ k = \text{numbers of words in a text} \]

Step 4: ngram

Unigram and Bigram are generally used in research as it includes features comprising of sets of two adjacent words. It was observed that unigram cannot capture phrases and multi-word expressions, effectively ignoring any word order dependence [25]. For example, words like ‘not excited’, ‘not satisfied’ clearly say that the sentiment is negative, but a unigram might fail to identify. The outcome of equation (4) considering all the words is shown in equation (5)

\[ \text{where, } \text{ngram}_{k} = k - (t - 1) \]
Part 2c: Topic Modeling 1-Latent Semantic Indexing (LSI) description

To identify K topics (K=5), the algorithm reads each text, \( d \), and goes through each word, \( w \), for topic \( T \), to compute

- \( P(T/d) \): proportion of words in texts \( d \) that are assigned to topic \( t \)
- \( P(w/T) \): proportion of assignment of words \( w \), from texts \( d \) to topic \( t \)

The complete expression is estimated as:

\[
p_{d} \cdot p_{w}(w|T)\]  \hspace{1cm} (9)

Repeating equation (9) several times generates steady state of appropriate word assignment to topics.

\[
p \left( \frac{T'}{d} \right) + p \left( \frac{w'}{T} \right) = p \left( \frac{T'}{w', d} \right)
\]

\[#\text{words in topic } T' + \#\text{words in } d \text{ belogs to topic } T' + \alpha \]  \hspace{1cm} (10)

Using equation (10), LSI correctly places the word \( w \) to the topics \( T' \) for which \( p \left( \frac{T'}{w', d} \right) \) is maximum.

III. OUTCOMES CENTERED ON PROPOSED MODEL

The outcomes are generated based on execution of proposed model for three steps (as mentioned above):

- Sentiment Analysis for 3 online stores (as stated in Part 2a)
  - Figures 2, 3 and 4 (a, b and c) portrays sentiments based on density.
- Classification of Sentiments into 8 polarities (as stated in Part 2b) – Table 1 divides the sentiments into 8 polarities. Tables 2, 3 and 4 identifies 5 relevant topics (words) which are used thoroughly in almost all the texts of 3 online retail stores.
and positive sentiments is (as mentioned in equation 7 and 8): 
positive = \{Joy, Surprise, Trust, Anticipation\} = \{663,129,332,11\} 
negative = \{Anger, Disgust, Fear, Sadness\} = \{11,10,33,231\}

C. Topic Modeling

The model identifies 5 most used positive and negative words used by consumers in each of the media are extracted. It was observed that consumers are using harsh words while mentioning their negative feedback, and normal terminologies were used while providing positive comment. It was also observed that none of the feedback (either positive and negative) were answered by companies.

V. CONCLUSION

Sentiment analysis is a field of study that examines people’s sentiments or emotions towards certain entities. This paper proposes a model to tackle a fundamental problems of sentiment analysis, sentiment polarity categorization and topic modeling. The site reviews from three websites (ComplaintBoard, Princena and Sitejabber) are selected as data source, wherein the model does web extracting and after initial preprocessing does sentiment analysis. A sentiment polarity categorization process has been proposed along with detailed descriptions of each step. Topic Modeling is conducted using Latent Semantic Indexing to identify the relevant topics of discussion in the study.

REFERENCES


IV. ANALYSIS AND CONCLUSION

The analysis of proposed model is divided into three parts, which are sentiment analysis, sentiment classification and topic modeling.

A. Sentiment Analysis

In the proposed model, sentiments of extracted texts are observed in terms of its density. “Density” signifies the intensity of sentiments in scale of 0 to 100, with positive and negative scale between 0 to 1. Souq.com – The negative intensity is higher in ComplaintBoard whereas, almost equally distributed for Sitejabber with and positive sentiments is (as mentioned in equation 7 and 8): 
positive = \{Joy, Surprise, Trust, Anticipation\} = \{663,129,332,11\} 
negative = \{Anger, Disgust, Fear, Sadness\} = \{11,10,33,231\}

C. Topic Modeling

The model identifies 5 most used positive and negative words used by consumers in each of the media are extracted. It was observed that consumers are using harsh words while mentioning their negative feedback, and normal terminologies were used while providing positive comment. It was also observed that none of the feedback (either positive and negative) were answered by companies.

V. CONCLUSION

Sentiment analysis is a field of study that examines people’s sentiments or emotions towards certain entities. This paper proposes a model to tackle a fundamental problems of sentiment analysis, sentiment polarity categorization and topic modeling. The site reviews from three websites (ComplaintBoard, Princena and Sitejabber) are selected as data source, wherein the model does web extracting and after initial preprocessing does sentiment analysis. A sentiment polarity categorization process has been proposed along with detailed descriptions of each step. Topic Modeling is conducted using Latent Semantic Indexing to identify the relevant topics of discussion in the study.

REFERENCES


AUTHORS PROFILE

Riktesh Srivastava is Ph.D. in ECE and Management from Indian Institute of Management, Ahmedabad (IIMA). Apart, he holds Masters in Electronics Engineering and Marketing Management, as well. Furthermore, completed prestigious certifications on Marketing Analytics and Electronic Commerce from Wharton School, University of Pennsylvania, USA and NTU, Singapore, respectively. He has written 3 books (OS, C++ Programming and RDBMS) and published more than 48 papers in International Journals and Conferences. He is also associated with several Scopus indexed journals. His area of interest includes Machine Learning, Big Data Analytics, Queuing Theory and currently indulged into Blockchain technology. Currently he is Faculty, School of Business at Skyline University College, Sharjah.

Dr. Mohammad Abu Faiz is currently working as Faculty at City University College, Ajman. Dr. Faiz completed his PhD in Business Management from Central University, Allahabad in 2005. His area of research is Electronic Business, Mobile Marketing, Business Ethics and Corporate Social Responsibility. His work has been featured in several national and international refereed Journals. Dr. Faiz is on the editorial board of various national and international journals of repute. Dr Faiz has published more than 30 papers in journals and conferences respectively.