**Saman Khan**

* **Research Paper # 01:**

**Link:** https://deliverypdf.ssrn.com/delivery.php?ID=074071112000081000029098000125082025056018064054034063029007091123094067075118081077004031019026053056124094121090115116023126118039014012004001002103092091125097107063029094003113101029072106090030011071004113080022020097086070087093029113124101125112&EXT=pdf&INDEX=TRUE

**Title:** PRODUCT SENTIMENT ANALYSIS FOR AMAZON REVIEWS

**Useful Info:**

* The enhancement in the field of e-commerce has led to a revolutionary change in the trading process. People‟s viewpoint have shifted from traditional commerce to ecommerce in the past years. In order to generate more traffic and increase in sales, merchants have enabled customers to share their opinion of the product. As A Result, Online Purchasing has grown, and that has led to Growth in Online Customer Reviews of Products.
* Our opinion and purchasing decision-making are affected by the experience of others and their feedback about products. We always ask others about their opinion to get the benefit from their experience; hence, the importance of reviews has grown.
* Sentiment analysis helps a customer to make their purchasing decision based on the experience of others. Further, it will help companies to improve their products by knowing customers’ opinions and needs.

**Methodology:**

* The proposed methodology of sentiment analysis for Amazon mobile phone reviews inludes phases from data collection until evaluating each classification model.
  1. First step is data pre-processing which inludes tokenization. This process relies on the space character to perform word separations. Then, all tokens were returned to their base or dictionary form by applying lemmatization process. Each review is labeled to positive, negative, or neutral based on its rating. Then, the dataset is split into testing and training.
  2. Bag-of-Words, Term frequency - Inverse document frequency, GloVe are being used for feature extractiion.
  3. Supervised machine learning classifier model, such as Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and so on are used for data training. After the training phase with the sentiment labels, the classifier can be used to predict the sentiment orientation of a sample on new data.
  4. The performance of the classification methods can be found by using Accuracy, F-Score, Crossentropy, Recall, and Precision. These parameters are helpful to evaluate the performance of supervised machine learning algorithms, based on the element from a matrix known as the confusion matrix or contingency table.

**Data Collection and Analysis:**

* An Amazon dataset extracted via Prompt Cloud is considered for sentiment analysis. The dataset concerns unlocked mobile phones and it was acquired in December 2016. It is a publicly available dataset from Kaggle.com. The Amazon reviews dataset consists of beyond 400,000 consumer reviews in the category of mobile phone. Particularly, it covers 413,840 reviews and 6 features, classified as follows: i) Information about mobile phone (Brand Name, Product Name, Price, Rating). Ii) Information about reviews (Reviews and Review Votes)

**Results & Discussions:**

* Python and its notebook, Jupyter, was used in this project alongside other supporting libraries to accomplish data cleansing, visualization, pre-processing, and machine learning modelling. Google Colab was used for faster implementation.
* After analysing the results of all the classification models used, the BERT model outperformed. On the other hand, it is observed that the BERT model’s success in classifying the neutral class achieved only 65%; this was expected, since a review with three-star rating does not mean the customer was absolutely balanced in their opinion between positive and negative. On the other hand, it achieved 97% in classifying positive reviews and 94% in classifying negative reviews.

**Conclusion:**

* Extracting information about a product from the review will help a customer to exploring more about the product and help them in decisionmaking.
* In this study, multiclass and binary classification for Amazon mobile phone using supervised machine learning algorithms: Logistic Regression, Naïve Bayes, Random Forest along with different feature extraction approaches is studied. Further, this project applied Bidirectional Long-Short Memory (Bi-LSTM) with GloVe embedding and joint-learned embedding.
* Moreover, Bidirectional Encoder Representations from Transformers (BERT) model was also applied. BERT model has achieved an excellent result in multiclass classification and binary classification, with accuracy of 94% and 98%, respectively. On the other hand, Bi-LSTM with joint-learned embedding also provides a very good result, with accuracy of 93% for multiclass classification and 97% for binary classification. Random Forest with word embedding (GloVe) outperforms other baseline models, LR and NB, with accuracy of 90% for multiclass classification and 94% for binary classification.

**Summary:**

As the world is drifting away from traditional means of trading, it is leaning more towards the revolutionary world of e-commerce. Due to this online purchasing has increased exponentially for which merchants are utilizing online feedbacks for product recognition. These opinions are helpful for other users to get a clear view if a product is worth their while or not. This is where the sentiment analysis comes in to determine your purchase decision based on the experience of others. The methodology required in this analysis contains; tokenization, GloVe, supervised machine learning classifier model, and for performance measurement; accuracy, F-score, Crossenthropy, recall, and precision. Datasets are needed for large scale AI models to help determine what your preference should be based on others feedbacks. As an example we can consider an amazon dataset taken from kaggle.com. It includes 400,000 plus reviews on mobile phones mainly their name, price, model etc and the information regarding reviews. Jupyter plus its supporting libraries are used for the cleansing, visualization and pre-processing of this project. After successful implementation neutral, positive, negative reviews got 65%, 97%, 94% respectively. In the given study, multiple AI algorithms such as logistic regression, naïve bayes, random forest and many more were implemented. Moreover, Bidirectional Encoder Representations from Transformers (BERT); which had an excellent result in multiclass classification and outperformed other models; was also applied and by the amalgamation of these concepts the desired results for sentiment analysis were achieved.

* **Research Paper # 02:** (Whatsapp wala)

**Title:** Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews

**Useful Info:**

* Opinions are statements that reflect people’s perception or sentiment. Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language helping in finding the mood of the customers about a purchasing of a particular product or topic.
* Data used in this paper is a combination of product reviews collected from Amazon.com, between July and September, 2018.
* The whole process includes summarization in three steps: (1) Product feature based, which is given by customer; (2) In each review, Identify expected features in each opinion sentence and (3) Finding out whether the feature/opinion is positive, negative or neutral and finally summary will be created.
* Sentiment analysis, also known as Opinion mining, is the study of sentiments that determines the judgement of people’s opinions, sentiments, evaluations, and emotions in relation to entities such as products, services, organizations, events, topics and their different attributes.

**Research Pzttern & Methodology:**

* The appropriate amount of Data used in this paper is arranged set of product reviews collected from amazon.com. From August to December 2018, in total, we collected over 500 sentiments of product reviews in which the products belong to 4 major categories: Mobiles, Computers, Flash drives and Electronics. These online reviews were posted by over 3.2 millions of customers (reviewers) towards 10,001 products.
* All subjective content was extracted for future analysis which consists of all sentiment sentences. A sentiment sentence is that which contains, at least, one positive or negative word.
* A process known as sentiment polarity categorization and POS has been proposed along with detailed descriptions of each step. These steps consist of pre-processing, pre-filtering, biasing, data accuracy etc. features which require the knowledge of machine learning.
* In sentiment classification, a POS tagger is most important because of the following two reasons: 1) Words like nouns and pronouns mostly do not contain any sentiment. So, it is able to filter out such words with the help of a POS tagger; 2) A POS tagger can also be useful in distinguish words that can be used in different parts of speech.
* The POS tagger used for this survey is a max-entropy POS tagger developed for the Penn Treebank Project. This tagger is able to provide 46 different tags which indicate that it can identify more detailed semantic roles than only 8.
* A lot of work in opinion mining and sentiments of customer reviews has been conducted to mine opinions in form of document, sentence and feature level sentiment analysis. For future preferences, Opinion Mining can be carried out on set of discovered feature expressions extracted from reviews become a most interesting research area. There is more innovative and effective techniques have to be invented which should overcome the current challenges faced by Opinion Mining and Sentiment Analysis.

**Summary:**

Sentiment analysis is one of the fastest spreading areas in computer science which uses a series of tolls and methods for detecting and extracting subjective information. An amazon report of July to September 2018 is used in this project to tackle sentiment polarity distribution. The whole process is summarized into Product feature based, which is given by customer, Identify expected features in each opinion sentence, finding out whether the feature/opinion is positive, negative or neutral. In the given project all subjective content was extracted for future analysis which consists of all sentiment sentences. A process known as sentiment polarity categorization and POS has been proposed which consist of pre-processing, pre-filtering, biasing, data accuracy etc. The POS tagger used for this survey is a max-entropy POS tagger developed for the Penn Treebank Project. Opinion Mining can be carried out on set of discovered feature expressions extracted from reviews to become a most interesting research area.