Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.

Extraction of Polarity from Textual Information based on Machine Learning Approach

Nitin B Raut[1], **Niveda M** [2], **Nivetha G**[3], **Parameshwari P** [4]

1, Assistant Professor Department of Computer Science and Engineering
KPR Institute of Engineering and Technology
[2,3,4] UG Students, Department of Computer Science and Engineering
KPR Institute of Engineering and Technology

ABSTRACT

Many of the major corporations use the public platform to advertise their products and services. As a result, the feedback given by their numerous customers is the foundation of their popularity among the general public. Manually categorizing the feedback into positive polarity or negative polarity can be a huge time consumable. The classification process is fully automated to improve results. Since reviews are inherently unstructured, they should be pre-processed and POS-labelled before being used to categorize human polarity. Term Frequencies are used to extract and weigh the terms that are mainly sentiment bearing (TF). The reviews are classified using supervised Machine Learning algorithms. The working of the algorithms for classification is evaluated using performance parameters such as Support terms, Term Frequency, Inverse document frequency, weight vector, and opinion power.

Keywords

POS Tags, Polarity Extraction, Term Frequencies, Support words, Opinion strength

Introduction

Polarity extraction, also known as opinion mining, is a method of assessing a person's emotions in relation to their behaviour using a set of terms. It is used to gain a simple understanding of the various forms of behaviours, thoughts, and emotions that people have, as well as their related qualities.It is extremely beneficial in the current situation because it offers information about every product based on various reviews, articles, and comments. Opinions and issues derived from social networking networks and online marketing outlets such as E-Commerce websites serve as a valuable source of knowledge for further research and better decision making. The level of each sentence decides whether it expresses a neutral, negative, or positive opinion.The sentiment classification at the Aspect level focuses on all expressions of sentiment found in a text as well as the aspect to which it relates. Sentiment Analysis can be used in a variety of ways.It can be broken down into two key areas.The Classifier is in charge of categorising texts into positive polarity, negative polarity, or neutral polarity. Unsupervised learning, unlike supervised learning, cannot be processed easily because no label data is needed. Since reviews are typically unstructured and can contain unnecessary information, they are subjected to advanced-processing, which involves eliminating extra symbols, punctuation, and numbers, among other things.The POS is done using pre-cleaning feedback to retrieve the functionality. The number matrix is generated from tagged reviews and used as an input to classifiers. The success indicators are used to assess the reviews' effectiveness. The contributors include:

- Text data is transformed into a matrix of numbers using various weighting methods such as Term Frequency (TF) and Inverse Document Frequency (TF – IDF) Weighting.
- For the classification algorithms, performance is measured using metrics such as support terms, weight vector, TF, IDF, and opinion power.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.

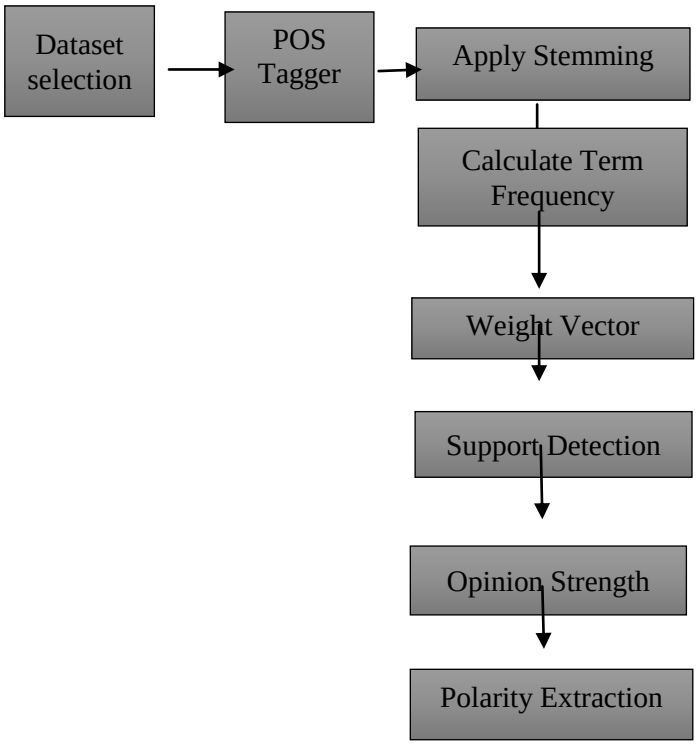
Literature Review

Table 1. Literature Review

Authors	Approach	Key Findings
Peter D. Turney	Unsupervised machine learning algorithm	They proposed a system known as Thumbs up or Thumbs down, which could be a simple and not a supervised learning algorithm for classifying input that has recommended or has not.
Pang et.al	Classifications and categorisation study	Hehave taken into account of the various aspect of classification of sentiments with respect to positive and negative sentiments based on the categorization study
Pang, Lee, Vaithiyanathan	Supervised machine learning algorithm	They have initiated the sentiment analysis process with the help of the well versed three types of supervised machine learning algorithm which uses the combinations of unigram bigram and n-gram.

Methods

Architecture Diagram



Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.

Figure 1. System Flow

Reviews:

The dataset includes reviews of the products and masterdata from various E-commerce websites that were extracted from the KDD (Knowledge Discovery Database), which contains product reviews from various domains such as clothes, cosmetics, accessories, kitchen appliances, mobile phones, DVDs, and books.Customers' product reviews are included in this dataset. The Knowledge Discovery Database provided the data for this dataset. This project's sample dataset comprises 2000 tests, 1700 of which are testing data and 300 of which are research data.

POS Tagging:

In text linguistics, POS tagging is the process of marking up a term using POS tagger software, also known as Parts of speech tagging, Grammatical tagging, or word-category clarification.in which the part of speech for each word is analysed based on both its meaning and context. The terms that add the broad and more precise sentiments can be chosen and considered for review by obtaining the POS tags of each and every word in the text. Nouns, verbs, adjectives, and adverbs are all essential components of the POS weighting system, which operates on the theory of stemming and marking, speeches are assigned higher weights than other POS tags. Some of the POS tags that express sentiments are described in a tabular format below, with the POS tags and their associated meanings.

NN RB RBR RBS	Noun Adverb Adverb, Comparative Adverb, Superlative
PRP IN	Personal Pronoun Preposition
VB VBP	Verb Verb, Present
JJ JJR	Adjective Adjective, Comparative
MD	Modal
FW	Foreign word
CD	Cardinal Number
TO	‘to’

Figure2. POS Tags

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.

The Dress looks pretty

Figure3. Example Review without POS Tagging

The Dress/**NN** looks/**VB** pretty/**JJ**

Figure4. Example Review with POS Tagging

Methodology

Suffix Stripping Algorithm

Suffix stripping algorithm is used to minimalize the use of derived words from their base words. This algorithm simply removes or strips the suffixes out of the specific words to which the application of this algorithm is possible.

For instance, the words ending with,

- "ed"
- "ing"
- "ly"
- "tion"

These ending words are removed from the base words and they are generalised with the help of the porter stemmer algorithm
For example, the words such as

- Computing
- Computation
- Computer

Gives the same meaning but their part of speech and their spellings are different. So, the main aim of the suffix stripping algorithm is to generalise these two or more words having the same meaning. So, the above mentioned three words are generalised to a common word "Compute"

Porter Stemmer

Stemming is the process which is often used in information retrieval techniques in search engines for filtering the redundant or the derived words from the document corpus (or sometimes derived) it reassembles the words to their word stem, base or root form. It helps in

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.

removing the extra words by generalizing the most common words into a common word.

Association Rule

- It is applied on the pre-processed dataset to find all frequent item sets.
- The frequently occurring combinations of words are obtained by calculating the support and confidence formulae.

Support words

The number of occurrences of a particular itemset to the total number of datasets is referred to as support

Support(s) = (A+B)/Total

Confidence

Confidence refers to the ratio of occurrences of support of A union B to the support of B

Confidence(c) = Support (A U B)/Support(B)

Weight Factor:

Weight factor is the multiplication of logarithmic values of term frequency and the corresponding inverse document frequency

Weight vector (wv) = TF*IDF

Opinion Strength

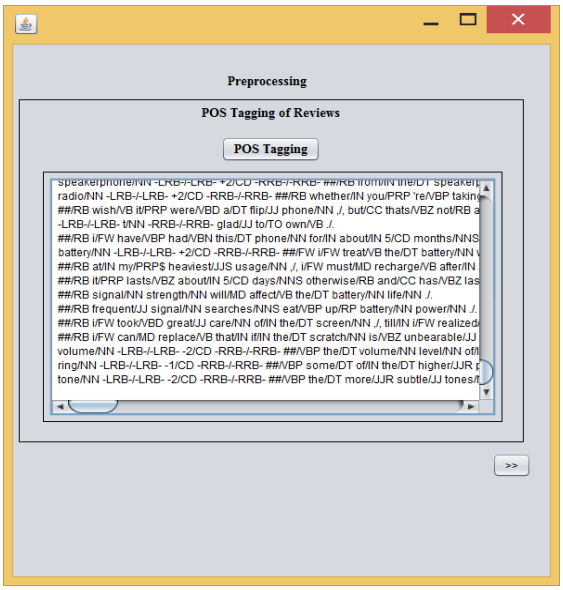
The opinion strength is calculated from the extracted feature words by using the stemming process. Opinion terms are crucial in deciding whether textual material is classified as positive or negative.
For this purposes Wordnet is used as the resource

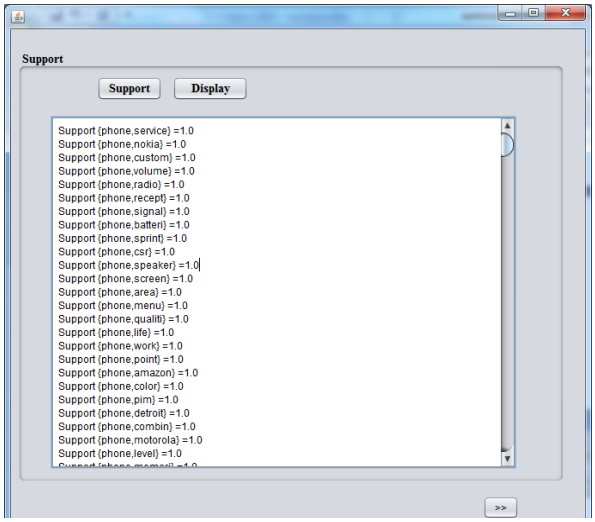
Classification

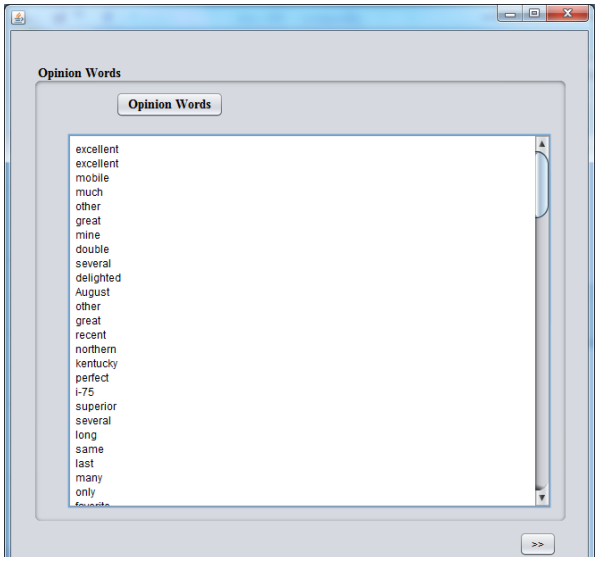
The opinionated words are then classified as positive and negative polarities with help of semantic orientation.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.

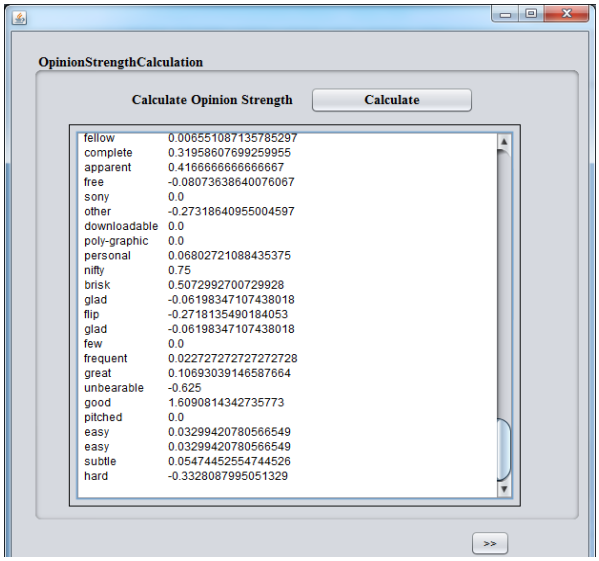
Data Analysis

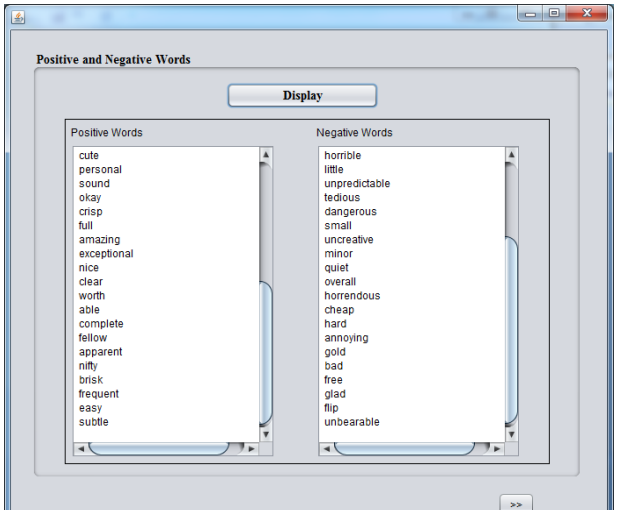






Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.

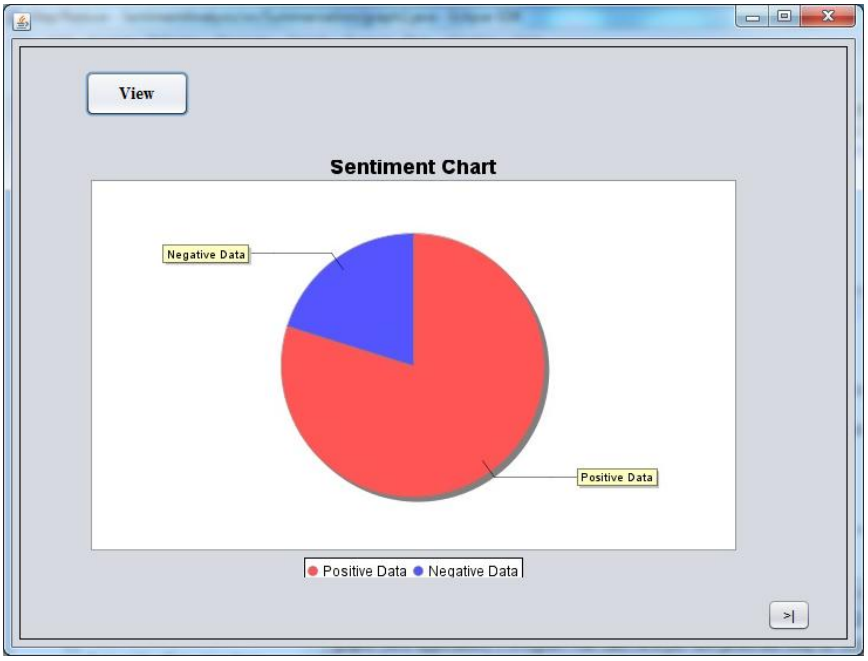




Results

The textual information which is given by the users are taken into account and thoseinformation are subjected to a various process of algorithms and analysis. Finally, a polarity pattern is extracted which is displayed below. It shows the 2 different aspects of the user which can be used for the improvement of the products.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5138 - 5146
Received 15 May 2021; Accepted 20 May 2021.



Conclusion

Polarity extraction normally considers only the sentiment of text or emotions considering both is a challenging factor. Here we are doing semantic of customers using the feedback data of product from ecommerce website by considering the text and emotion in the review. We are also going with a recommendation system that recommends the customers with product of their area of interest based on the factors like age, gender, work-sector etc.,

Future Studies

We will apply the different forms of clustering algorithms to a broader range of polarity analysis tasks in the future. More complex polarity change patterns, such as transitional, subjunctive, and sentiment-inconsistent sentences, may also be considered. This will create a huge impact and a greater advantage in the future by which we can include and aggregate various reviews from two or more sites and can easily implement the fastest clustering algorithm to reduce more computational time for the companies which are aiming the customers satisfaction factors and can highly engage people to buy their products based on the reputations they gain on their social media platforms as well as they can improve their standard of marketing strategy.

References

[1] S Muthukumaran , P Suresh, AVA Mary., “Sentiment Analysis for Online Product Reviews using NLP Techniques and Statistical Methods”

[2] S Shubha , P Suresh.., “An efficient Machine Learning Bayes Sentiment Classification method based on review comments”

[3] SA Jadhav, DVLN Somayajulu, SN Bhattu, RBV Subramanyam, P Suresh.., “Topic dependent cross-word spelling corrections for web sentiment analysis”

[4] “Technology factors of online shopping and their effect on attitude with special regard to the student community,” RR Kumar, DR. P. Thangaraj, Siva Sangari M., Devipriya A., M.Salomi Samsudeen.

[5] C.G. Akcora, M.A. Demirbas,M.A. Bayir, M. Ferhatosmanoglu, H.:Public opinion breakpoint identification.

[6] Ahn, J.H., Baek, H.M., and Oh, S.W.: The Impact of Tweets on Movie Sales: The Period When Tweets Are Posted. J. ETRI.

[7]A. Boutet, H. Kim, and E. Yoneki: What Do Your Twitter Messages Contain? I know who you voted for in the UK election of 2010.

[8] Diakopoulos, N., and Shamma, D.A.: Using Aggregated Twitter Sentiment to Characterize Debate Results.

[9]ARSA stands for "A Sentiment-aware Model for Predicting Sales Performance Using Blogs."Liu, Y., Huang, X., An, A., and Yu, X.

[10] Detection Issue and Prediction Analysis on Big Data: Trends in Electronics and Telecommunications, J. Lee, C.H., Hur, J., Oh, H.J., Kim, H.J., Ryu, P.M., Kim, H.K.