# Sentimental analysis on Amazon Reviews

K.Varuna Krishna
*CSE*
*IIIT Sri City*
*Andhra Pradesh, India*
*Email: varunakrishna.k19@iiits.in*

G.Vikram
*CSE*
*IIIT Sri City*
*Andhra Pradesh, India*
*Email: vikram.g19@iiits.in*

k.venkatesh
*CSE*
*IIIT Sri City*
*Andhra Pradesh, India*
*Email: satyadurgavekatesh.k19@iiits.in*

D.karthik
*CSE*
*IIIT Sri City*
*Andhra Pradesh, India*
*Email: satyakarthik.d19@iiits.in*

B.koushik
*CSE*
*IIIT Sri City*
*Andhra Pradesh, India*
*Email: koushik.b19@iiits.in*

***Abstract*—Sentimental analysis is the mainly needed now a days.Finding sentiment of the data will be useful for different purposes.We used sentimental analysis task on amazon reviews. our contributions are *i)* collected a new dataset, *ii)* POS tagger, *iii)* Rule based lemmatizer, *iv)* our proposed pipeline.**

## 1. Introduction

Social media has become increasingly important to our daily lives, and many people are writing blogs, reviews over internet. As Now a days there is a huge amount of data around us,figuring out information from them is very hard.One such information is sentiment of data.Finding sentiment of reviews from amazon is much needed now a days.It will be helpful to many Business analyst's, Risk analyst's of product based companies to figure out how their product is reached in markets.It will be a tedious task to go through each and every review.Our work orchestrates the work of finding sentiment's on amazon reviews. we developed a new pipeline and created our own data set. Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. With the sentiment information you can manage conversation flow or perform post-call analysis. For example, if the user sentiment is negative you can create a flow to hand over a conversation to a human agent.

## 2. State of the art

[1] This research paper provided the sentimental analysis various opinions on smart phones dividing them as positive, negative and neutral behaviour. The data for this research is collected of online product reviews from Amazon.com. 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 .

[2] This paper focus on the technique of providing automatic feedback on the basis of data collected from twitter.Here they mainly analyze the data for mobile phones.We collected the data's from twitter page using the crawler Twitter4J API. After that we stored the crawled data to a standard database. Then we preprocess it by eliminating the stop words. Then the data's from the database are classified using POS tagging approach and the tag set is being created. The data's from the tag set are fed to the SVM algorithm.. The experiments have shown 80% accuracy in the sentimental analysis

[3] In this paper, they design a framework for sentiment analysis with opinion mining for the case of hotel customer feedback.The proposed framework is termed sentiment polarity that automatically prepares a sentiment dataset for training and testing to extract unbiased opinions of hotel services from reviews. They selected the OpinRank dataset because it contains unlabelled reviews that gave The system is expected to determine sentiments the way human beings do and labelled datasets are normally used for the system to learn automatically. The proposed framework tries to make sure that sentences are correctly labelled such that false information is not fed into the system,the flexibility for a custom experimentation.

[4] In this paper sentimental analysis is performed on the data extracted from Twitter and Stock Twits. The data is analyzed to compute the mood of user's comment. These comments are categorized into four category which are

happy, up, down and rejected. The polarity index along with market data is supplied to an artificial neural network to predict the results.

[5] This paper shows a analysis of sentiments to get the opinion of people if they are positivity during this situation or not. The paper is using the technique of polarity to know the opinion is positive, negative, or nonpartisan . Three main keywords ”COVID”, ”Corona virus” and ”COVID-19” are used to check the polarity

[6] In this the target is to deal with the problem of Review System, an utmost important part of any organizational CRM.. Data inflow in this project is through the twitter API supplying live stream of tweets.. Finally we would set a stage to provide insights into our future work on sentiment analysis and using this Smart Agent Analysis on existing CRM in order to improve their existing Feedback structure. Answers to these questions are provided by statistical analysis on keyword.

[7], The main aim of this paper is to apply sentimental analysis of Indian languages on Twitter data. For this the proposed model first scrape the tweets from Twitter by using Twitter APIs, then later by using text blob, the customer reviews are given different sentiment scores and classify them as positive, negative, or neutral.The classification techniques used are Support Vector Machine (SVM) and K Nearest Neighbor (KNN). It achieved an accuracy of 98%. The proposed model deals with different Indian regional languages (i.e., English, Kannada, Malayalam, Tamil, Telugu, and Hindi) simultaneously.

[8],The main aim of this paper is to apply sentimental analysis on Twitter data with respect to general elections. Twitter APIs are used for the collection of tweets. In this paper, R is used for the acquisition, pre- processing, analyzing the tweets, then sentiment analysis is performed based on the different approaches like lexicon and NRC dictionary based approches are used.The actual result of the election obtained in May 2019 was in full compliance with the results obtained during this experiment from Jan 2019 to March 2019. This shows that the approach applied in this paper for performing sentiment analysis was apt.

[9],In this paper, a new model to predict home location for Twitter users based on sentiment analysis (Pre-HLSA) is proposed. It predicts the users’ home location using only their tweets, by analyzing some of the tweet’s features.WordNet and sentiment analysis are applied for tweets to identify and extract their sentiments and polarity. The experimental results show a promising performance compared to the previous methods in terms of accuracy, mean and median performance measures. It achieves up to 85% accuracy, 223km mean, and 96km median.

[10],The objective of the paper is to understand the sentiments of Indian citizens towards the COVID-19 vaccine.For this study, they used the python library Textblob to process the textual data. Textblob, using NLP (Natural Language processing) and advanced machine learning principles, analyze every word in the documents presented in the corpus and define the overall sentiments being projected as positive, negative, or neutral. The results show that only 35% of social media posts (n=73,760) about COVID-19 vaccines were in a positive tone.

In paper [11] the author discusses about the systems are designed to retrieve information using twitter data and then classify them based on the semantics of knowledge contained. Authors Lokmanyathilak Govindan Sankar Selvan and Teng-Sheng Moh have developed the framework.

in paper [12] which makes use of real-time Twitter data stream, that are cleaned and analyzed and then fast feedback is acquired through opinion mining.

Paper [13] deals with the opinion extracted and collected from the popular social media platform named Twitter. For Comparison of market status of two enterprises they copy the two dictionary files one for positive and other for negative words from the repository in the backend as those files are used for analyzing and scoring terms from tweets.

The paper [14] mainly convey about the sentimental analysis of tweets using R language which is helpful for collecting the sentimental information in the form of either positive score, negative score or someplace in between them. Then they execute the analysis of tweets that are in size of TBs which means big data using R language and Rhadoop Connector.

In paper [15] author has considered a way of advancing the present sarcasm detection algorithms by including improved pre-processing and text mining techniques like emoji and slang detection. To analyze bulky live data streams, three considered key features are density distribution, negativity and influence.

To do this in paper [16] authors Ming Hao, Christian Rohrdantz and others have considered Pixel cell-based sentiment calendar where every opinion is represented in form of cell. The cell color is the sentiment value, i.e, green for positive,gray for natural and red for negative.

In paper [17] authors Sonia Anastasia and Indra Budi have used the Sentiment analysis tools: R and Rapidminer. Data is collected by crawling through Twitter API with R. Then, Rapidminer is used in data pre-processing, classification, for classifiers performance evaluation and Net Sentiment Score is calculated which associates with customer satisfaction using classification results.

In paper [18] tweets are collected and all unimportant words are deleted from the tweet collection to support the classification process, then tweets are filtered by means of a Bayes Naive classifier, which was earlier trained and whose intention is to select messages that represent news about fresh cyber-attacks and malware. In this paper exemplary malware activities presently circulating on the network have been identified.

In paper [19] authors give two easy yet effective ensemble beginners for Twitter data sentiment classification. This paper presents a good solution to effectively classify the tweets related with product A, one can use tweets related with other products in the same class of product A to train a classification model. This recommends the homogeneity of tweets from related products in the context of sentiment classification.

In paper [20] author give details about usage of StanfordcoreNLP libraries and twitter4j libraries to construct an application that can obtain data in the form of tweets and perform the sentimental analysis to display positive or negative tweet on any particular topic using its associated hash tag. Authors M.Trupthi and others focus on the short sentences and entity level sentiment analysis in paper [21].

## 3. Dataset

We collected a data set consisting of 19.9K reviews. These reviews were collected from amazon website. We first collected 80 url's of different products in amazon. then we used a crawler to fetch all the reviews.
We have made sure that every review is in English language. we removed any review which is not in English and we also removed reviews which only has emoji's.

### 3.1. Annotation

we annotated the reviews, we collected, into two classes. These classes are positive and negative.We used an xlsx UI for annotation. Each annotator gets a query and he decides the class of the review.The data analytics is in table 1

| positive | negative |
|---|---|
| 12.6 K | 7.3 K |

TABLE 1. Data Analytics

## 4. Proposed System

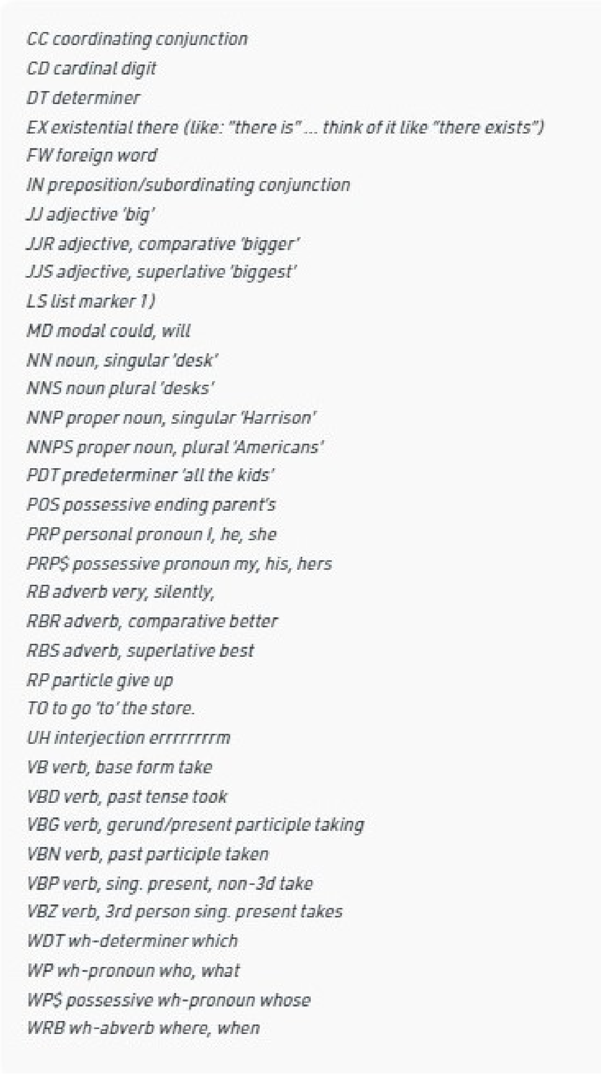
In this work we have build our own POS tagger, our own rule based lemmatizer , and a proposed Model.

### 4.1. POS Tagger

If we talk about Part-of-Speech (PoS) tagging, then it may be defined as the process of assigning one of the parts of speech to the given word. It is generally called POS tagging. In simple words, we can say that POS tagging is a task of labelling each word in a sentence with its appropriate part of speech. We already know that parts of speech include nouns, verb, adverbs, adjectives, pronouns, conjunction and their sub-categories.
A set of all POS tags used in a corpus is called a tagset. Tagsets for different languages are typically different. They can be completely different for unrelated languages and very similar for similar languages, but this is not always the rule. Tagsets can also go to a different level of detail. Basic tagsets may only include tags for the most common parts of speech (N for noun, V for verb, A for adjective etc.). It is, however, more common to go into more detail and distinguish between nouns in singular and plural, verbal conjugations, tenses, aspect, voice and much more
POS tags make it possible for automatic text processing tools to take into account which part of speech each word is. This facilitates the use of linguistic criteria in addition to statistics.
For languages where the same word can have different parts of speech, e.g. work in English, POS tags are used to distinguish between the occurrences of the word when used as a noun or verb.
POS tags are also used to search for examples of grammatical or lexical patterns without specifying a concrete word, e.g. to find examples of any plural noun not preceded by an article. Or both of the above can be combined, e.g. find the word help used as a noun followed by any verb in the past tense
Sometimes a word on its own can give useful clues. For example, 'the' is a determiner. Prefix 'un-' suggests an adjective, such as 'unfathomable'. Suffix '-ly' suggests adverb, such as 'importantly'. Capitalization can suggest proper noun, such as, 'Meridian'. Word shapes are also useful, such as '35-year' that's an adjective.
Part of Speech (hereby referred to as POS) Tags are useful for building parse trees, which are used in building NERs (most named entities are Nouns) and extracting relations between words. POS Tagging is also essential for building lemmatizers which are used to reduce a word to its root form
A word can be tagged based on the neighbouring words and the possible tags that those words can have. Word probabilities also play a part in selecting the right tag to resolve ambiguity. For example, 'man' is rarely used as a verb and mostly used as a noun
There are eight classes that are generally accepted (for English). In alphabetical listing:
Adjective: beautiful, green, awesome...
Prepositon: to, with, in...
Adverb: hardly, above, soon...
Articles: a, the, an...
Conjunction: and, but, yet...
Noun: cat, ape, apple...
Pronoun: it, he, you...
Verb (including auxiliary): to be, working, stood

we created out own POS tagger. we randomly picked 2K samples from our data and annotated them. Each word in a

sentence get a tag. we used these rules, refer figure 1, for annotating. for Double checking we also used NLTK POS Tagger to verify, one problem in this is NLTK POS Taggers don't have that many tags we annotated.
After the tagging is done, we used this a training data and trained a decision tree and using that model we tagged other data.



CC coordinating conjunction
CD cardinal digit
DT determiner
EX existential there (like: "there is" ... think of it like "there exists")
FW foreign word
IN preposition/subordinating conjunction
JJ adjective 'big'
JJR adjective, comparative 'bigger'
JJS adjective, superlative 'biggest'
LS list marker 1)
MD modal could, will
NN noun, singular 'desk'
NNS noun plural 'desks'
NNP proper noun, singular 'Harrison'
NNPS proper noun, plural 'Americans'
PDT predeterminer 'all the kids'
POS possessive ending parent's
PRP personal pronoun I, he, she
PRP$ possessive pronoun my, his, hers
RB adverb very, silently,
RBR adverb, comparative better
RBS adverb, superlative best
RP particle give up
TO to go 'to' the store.
UH interjection errrrrrrrm
VB verb, base form take
VBD verb, past tense took
VBG verb, gerund/present participle taking
VBN verb, past participle taken
VBP verb, sing. present, non-3d take
VBZ verb, 3rd person sing. present takes
WDT wh-determiner which
WP wh-pronoun who, what
WP$ possessive wh-pronoun whose
WRB wh-abverb where, when

Figure 1. POS Tags used

### 4.2. Rule Based Lemmatizer

Lemmatization is the process of finding the normalized form of a word. It is the same as looking for a transformation to apply on a word to get its normalized form.
So a lemma is the base form of a word — this means that any variation related to time or quantity is removed. For example, in nouns, plurals (girls, boys, corpora) get reduced to its singular form (girl, boy, corpus); and in verbs, time/participle variants (ate, brought, chatting) are back to present tense (to eat, to bring, to chat).
In some cases, lemmatization can include removing gender variation (doctress → doctor), although it is very uncommon in English, since the language moved towards gender neutrality — however, it can be done in some specific cases (such as focusing onto a specific gender. Ex.: 'bull' → 'cow', 'rooster' → 'chicken').
Even though lemmatization might not seem as useful at first, it is a powerful tool for text normalization, since it allows normalization to occur in a more syntactical manner (verbs continue being verbs, nouns continue being nouns and so on) than stemming
We wrote the rules for doing lemmatisation. The process of this is : first we get a word to get it's lemma , then we find the POS tag using the tagger we build. Now, we got word and it's POS tag, we send these two values to our rule based lemmatizer. based on the rules it will do necessary changed to the word and produce it as lemma.

Some sample rule will be like:

```
    if Tag=="VBG" and Word="{WORD}ing"
      then lemma= WORD
```

example: lemma for running is run.

```
    if Tag=="NNS" and Word="{WORD}s"
      then lemma= WORD
```

example: lemma for boys is boy.

```
    if Tag=="VBN" and Word="{WORD}en"
      then lemma= WORD
```

example: lemma for taken is take.

```
    if Tag=="NNS" and Word="{WORD}ies"
      then lemma= WORD
```

example: lemma for thieves is thief.

```
    if Tag=="NNS" and Word="{WORD}s"
      then lemma= WORD
```

example: lemma for boys is boy.

### 4.3. Proposed Model

We have developed a pipeline for this task.
it goes in this fashion :
First a review will be sent as an input to the system, Then this sentence will be cleaned( we remove URL's, stop words,words with digits special characters,etc), this cleaned sentence is sent for Lemmatizer. After the sentence all words in the sentence will be clean words and will be lemma's. Now this sentence is passes to POS tagger and it will tag every word. Now from these tagged words we pick only Nouns,Verbs,Adverbs,Adjectives and remove the words with remaining tags. The reason for this removal is Other tags does not add much meaning to the sentence. Now this new query is passed to logistic regression for classification.

Another experiment we have done is same as the previous experiment but instead of passing modified review to Logistic regression we pass the clean review for classification.

## 5. Results

We have experimented two methods by passing : *i)* Modified review, modified using POS Tagger, *ii)* Original review. we can see the results from Table2 have increased when using modified review or refined review over original review.
we can say the reason behind : A lot of reviews are long and has unnecessary words even after cleaning. Passing important words from the sentence will make the model to understand the inner meaning in a more clear way.

| Model | Accuracy | F1-Score |
|---|---|---|
| using Modified review | 61 | 0.60 |
| using Original review | 63 | 0.70 |

TABLE 2. RESULTS OF OUR EXPERIMENTS

## 6. Conclusion & Future Work

we have developed a lemmatizer, which is light weight. here we are not using any model to train , we only use rule to find so it will predict in no-time.
Our work shows that passing the refined review gives better result.
In future work, we will develop a new word embedding space, specially to understand review data.

## References

[1] P. Pandey, N. Soni *et al.*, “Sentiment analysis on customer feedback data: Amazon product reviews,” in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. IEEE, 2019, pp. 320–322.

[2] M. P. Anto, M. Antony, K. Muhsina, N. Johny, V. James, and A. Wilson, “Product rating using sentiment analysis,” in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, 2016, pp. 3458–3462.

[3] K. Zvarevashe and O. O. Olugbara, “A framework for sentiment analysis with opinion mining of hotel reviews,” in *2018 Conference on information communications technology and society (ICTAS)*. IEEE, 2018, pp. 1–4.

[4] S. K. Khatri and A. Srivastava, “Using sentimental analysis in prediction of stock market investment,” in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2016, pp. 566–569.

[5] S. Raheja and A. Asthana, “Sentimental analysis of twitter comments on covid-19,” in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021, pp. 704–708.

[6] P. Sharma, T. Choudhury, A. S. Sabitha, and G. Raj, “Smart sentimental agent analysis through live streaming data,” in *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*. IEEE, 2018, pp. 399–402.

[7] K. Rakshitha, H. Ramalingam, M. Pavithra, H. Advi, and M. Hegde, “Sentimental analysis of indian regional languages on social media,” *Global Transitions Proceedings*, 2021.

[8] A. Sharma and U. Ghose, “Sentimental analysis of twitter data with respect to general elections in india,” *Procedia Computer Science*, vol. 173, pp. 325–334, 2020.

[9] A. Mostafa, W. Gad, T. Abdelkader, and N. Badr, “Pre-hlsa: Predicting home location for twitter users based on sentimental analysis,” *Ain Shams Engineering Journal*, 2021.

[10] S. Praveen, R. Ittamalla, and G. Deepak, “Analyzing the attitude of indian citizens towards covid-19 vaccine–a text analytics study,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 2, pp. 595–599, 2021.

[11] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, “Precise tweet classification and sentiment analysis,” in *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*. IEEE, 2013, pp. 461–466.

[12] L. G. S. Selvan and T.-S. Moh, “A framework for fast-feedback opinion mining on twitter data streams,” in *2015 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2015, pp. 314–318.

[13] A. O. Egorova, N. S. Andryashina, and V. P. Kuznetsov, “Methodology of formation and realization of competitive strategy of machine building enterprises,” 2016.

[14] S. Kumar, P. Singh, and S. Rani, “Sentimental analysis of social media using r language and hadoop: Rhadoop,” in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2016, pp. 207–213.

[15] A. G. Prasad, S. Sanjana, S. M. Bhat, and B. Harish, “Sentiment analysis for sarcasm detection on streaming short text data,” in *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*. IEEE, 2017, pp. 1–5.

[16] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L.-E. Haug, and M.-C. Hsu, “Visual sentiment analysis on twitter data streams,” in *2011 IEEE conference on visual analytics science and technology (VAST)*. IEEE, 2011, pp. 277–278.

[17] S. Anastasia and I. Budi, “Twitter sentiment analysis of online transportation service providers,” in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2016, pp. 359–365.

[18] F. Concone, A. De Paola, G. L. Re, and M. Morana, “Twitter analysis for real-time malware discovery,” in *2017 AEIT International Annual Conference*. IEEE, 2017, pp. 1–6.

[19] H. S. Kisan, H. A. Kisan, and A. P. Suresh, “Collective intelligence & sentimental analysis of twitter data by using standfordnlp libraries with software as a service (saas),” in *2016 IEEE international conference on computational intelligence and computing research (ICCIC)*. IEEE, 2016, pp. 1–4.

[20] M. Trupthi, S. Pabboju, and G. Narasimha, “Sentiment analysis on twitter using streaming api,” in *2017 IEEE 7th International Advance Computing Conference (IACC)*. IEEE, 2017, pp. 915–919.

[21] K. Roy, D. Kohli, R. K. S. Kumar, R. Sahgal, and W.-B. Yu, “Sentiment analysis of twitter data for demonetization in india–a text mining approach,” *Issues in Information Systems*, vol. 18, no. 4, pp. 9–15, 2017.