Pooja Khanna, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 11, Issue 5, (Series-III) May 2021, pp. 33-42

RESEARCH ARTICLE OPEN ACCESS

Prioritization of Customer Reviews via Efficient Sentiment Analysis Technique

Pooja Khanna*, Sachin Kumar**, Darpan Khanna***, Pragya****
**(Department of Computer Science, Amity University Uttar Pradesh, Lucknow Campus*
*** (Department of Electronics & Comm. Engineering, Amity University Uttar Pradesh, Lucknow Campus*
****(Department of Computer Science, Amity University Uttar Pradesh, Lucknow Campus*
*****(Department of Chemistry, MVPG College, Lucknow*

ABSTRACT
With integration of disruptive technologies, i.e. AI, IoT, Space colonization, 3D printing, Blockchain, Robotics, Autonomous vehicles, and Virtual reality, sectors such as retail, marketing, corporate, finance, product development and service based all have become vastly influenced by online support systems they offer, reachability and online ratings. Customer feedback has become utmost important for the potential growth. Reviews and feedback help a lot in understanding market trends to attract customers. Owing to volume and diversity of reviews for a product - managing, segregation and prioritization of the feedbacks manually is a very tedious task especially when there are reviews and feedbacks occurring in millions. This generates a potential requirement of management of these reviews and feedbacks in an efficient manner considering commercial aspect attached to it. An autonomous and self-sustained computation is required for managing the task Work presented is an effort to make this task easy, model proposed takes into consideration all possible reviews from social networking sites to Blogs, for the study database of reviews was picked from Kaggle, a well-known online repository. Model first segregates them with positive and negative labels, assigns weight and then displays them according to the priority of the review, as in what updates have to be brought first into the product as compared to the others. Model proposed is at par with existing recommending systems with an advantage of assigning priorities, targeted recommendations with prioritized display of products can churn huge profits.
Keywords – Customer review, sentiment analysis, tokens, tweets, batch size.

Date of Submission: 01-05-2021 Date of Acceptance: 15-05-2021

I. INTRODUCTION

With the variety and options increasing for every product and services and also the in-product options provided by a variety of different organizations, ranging from small local companies, serving over a small area to the largest of the multi - national corporations having foothold in every country they have possibly covered. With such cutting-edge competition in the market, every organization wants to set its monopoly in the market, in order to maximize their profit over the area and also to win the customers for a considerable period of time till another competitor with equal power enters the market. To maintain the monopoly or even to gain more foothold in the market, even the small organizations need to improve themselves constantly and continuously. Gaining on improvement is possible in the best way by taking reviews and grievances from their own customers and the customers using similar products from other organizations. For this a grievance system has to be in place and active, present and regularly maintained for the improvement of the products and services provided by a particular company irrespective of the fact that whether the company is newly established or century old. Present times grievance system has seen a complete makeover with the help of technologies, such as incorporating Natural Language Processing for product evaluation, for a better review system and minimizing the human tendency of committing error. Feedbacks and reviews from the customers, whether good or bad work very positively for the growth of organizations. Feedback system has been utilized via a wide variety of ways and various methods being applied and incorporated. Sentiment analysis is one of the potential techniques for review analysis and very useful method to differentiate between a genuine review or feedback and the spams flooding the review lists. The grievance and review system is a heterogeneous system, as it not only involves the developer to do the technical part in the sentiment analysis but also the human resource personnel and the customer welfare department of the organization to be involved in the system. These techniques are

important to draw conclusions as to how the feedback must be utilized to make the product or the service better, this is something which has to be looked after by these departments and not by the technical personnel's. The technical department however are better equipped to analyze the outputs given by the smart models and systems which are now being extensively applied to various review / feedback datasets for the various companies and organizations. Models are generally managed and maintained by a team of data scientists. [1-5]

II. MOTIVATION

The Natural Language Processing (NLP) is a functional block that empowers PCs to understand, interpret and control human language. NLP draws from various requests, including programming designing and computational semantics, to its greatest advantage to fill the gap between human correspondence and PC understanding. The progression of NLP applications across the world has progressed with voice quality, pronunciation and pitch variations from person to person and even geographical locations define how people speak, systems try usually to anticipate that individuals should "talk" to them in a programming language that is definite, unambiguous and significantly sorted out, or through a set number of evidently verbalized voice bearings. Human talk, regardless of language isn't continually accurate - it is oftentimes faulty and the semantic structure can depend upon various confusing elements, including slang, neighborhood vernaculars and social setting.

As a human, you may talk and write in English, Spanish or Chinese. In any case, a PC's local language – known as machine code or machine language – is to a great extent unfathomable to the vast majority. At your gadget's most reduced level interaction happens not with words yet through a great many zeros and ones that produce consistent activities. To be sure, developers utilized punch cards to speak with the primary PCs 70 years prior. This manual and laborious procedure was comprehended by a moderately modest number of individuals. Presently you can say, "Alexa, I like this tune," and a gadget playing music in your home will bring down the volume and answer, "alright. Rating spared," in a humanlike voice. At that point it adjusts its calculation to play that tune according to your needs. Sentence structure and semantic assessment are two guideline systems used with ordinary language taking care of. Semantic structure is the course of action of words in a sentence to look good. NLP uses phonetic structure to overview noteworthiness from a language reliant on syntactic standards. Language structure systems used fuse parsing (semantic examination for a sentence), word division (which isolates an immense piece of substance to units), sentence breaking (which spots sentence confines in enormous works), morphological division (which segments words into social affairs) and stemming (which allotments words with articulation in them to root structures).

Semantics includes the utilization and importance behind words. NLP applies calculations to comprehend the importance and structure of sentences. Systems that NLP utilizes with semantics incorporate word sense disambiguation (which infers importance of a word dependent on setting), named substance acknowledgment (which decides words that can be sorted into gatherings), and characteristic language age (which will utilize a database to decide semantics behind words).

Current approaches to manage NLP rely upon significant learning, a kind of AI that takes a gander at and uses plans in data to improve a program's understanding. Earlier approaches to manage NLP incorporated an extra standards-based methodology, where less unpredictable AI counts were resolved what words and articulations to look for in substance and given unequivocal responses when those articulations appeared. However, significant learning is an undeniably versatile, natural strategy in which estimations make sense of how to perceive speakers' objective from various models, basically like how a youth would learn human language.

Notion Analysis is the automated methodology that uses AI to perceive positive, negative and fair suppositions from content. Idea assessment is commonly used for getting bits of information from electronic life comments, diagram responses, and thing reviews, and choosing data driven decisions. [6-9]

In this present reality where we create 2.5 quintillion bytes of information consistently, opinion investigation has become a key instrument for comprehending that information. An important for data analytics can be sentiment analysis. Sentiment Analysis can be applied at various degrees of extension:

1. Report level feeling investigation gets the assumption of a total record or section.
2. Sentence level feeling investigation acquires the estimation of a solitary sentence.
3. Sub-sentence level feeling investigation acquires the estimation of sub-articulations inside a sentence.

Types of Sentiment Analysis:

1. Fine grained sentiment analysis
2. Emotion detection
3. Aspect based sentiment analysis.
4. Intent analysis

5. Multilingual sentiment analysis

Model proposes to employ the aspect-based sentiment analysis for customer feedback for the target organizations product to improve their products and services for a better customer experience. [8-11]

III. METHODOLOGY

Sentiment analysis as a technique is a backbone of most of the online recommendations systems, applications include social media monitoring, customer support management, and analyzing customer feedback. In the background of sentiment analysis, advanced AI algorithms apply language deconstruction techniques, like tokenization, part-of-speech tagging, parsing, and lemmatization to break down and make sense of text. Only then can machine learning software classify unstructured text by emotion and opinion. All this is made from the merging and already present online public forums also called as web forums or social media outlets. These discussions go about as an asset loaded with rich content-based information fit to be mined. Estimation examination is an idea that manages individuals' assessment, feelings, frames of mind towards a specific field or on the other hand theme. This paper centres around the functions of an AI based model that screens popular assessment on inclining points on the social media stage, Twitter, to land at the enthusiastic part of the assessment. This is accomplished by the NLP approach that digs the information for enthusiastic prompts in view of predefined catchphrases and learns the extremity of the general visibility on the subject. The mass expression are very useful in providing a diversified view over any topic in the form of mass expressions. These diversified web forums provide bipolar views over the topic being discussed upon. This in turn helps the people who don't know anything about a particular topic to gain handsome knowledge about in after going through the forums present online.

One of the most common online web-based forum is twitter. Twitter is the world's driving assistance for news and person to person communication on which clients post and associate with messages or remarks known as 'Tweets'. Despite the fact that unregistered clients can just view tweets, enrolled clients can likewise post their very own tweets and share different tweets, this procedure is alluded to as 'Retweeting'. With its adaptable and indulgent posting behaviour, Twitter has gotten home to individuals from varying backgrounds that post and offer tweets. Twitter has additionally been an online ordinary for warmed discussions and has led the pack on web based breaking news. Explicit subjects or current issues are ordered furthermore, gathered under the 'Hashtag'. This way twitter provides a lucid and also a diversified platform for learning and also participating in healthy discussion.

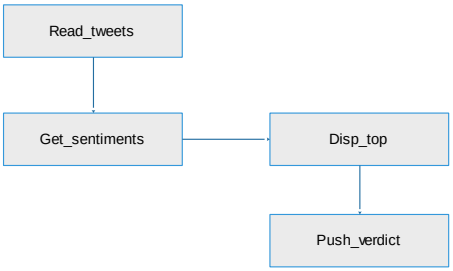


Fig.1: Flow chart depicts how the web based forums push the final verdict about the discussions.

Flowchart shows that algorithm reads all the tweets about the topics and observes all the sentiments in each. With each tweet divided according to the sentiments, the top ones are taken into account and then the model displays the final verdict regarding the discussion. This is a way in which a model can be made smart enough in its own ways thereby, reducing human intervention and interference in sentiment detection and analysis and hence, increasing the use of artificially intelligent models running on their own systematic algorithms to give more accurate, more efficient, process centric and taking less processing power of the automated systems. Another way in which the models are capable of working and also are brought to use is to displaying a particular bunch of output according to the needs of the data scientist. [11-14]

Aspect Based Sentiment Analysis (ABSA) gives further understanding into the investigation of online networking. Understanding client feeling about various parts of items, administrations or approaches can be utilized for improving and enhancing in a successful way. The standard pipeline of viewpoint comprises of three stages: viewpoint class location, Opinion Target Extraction (OTE) and slant extremity arrangement. In this article, we propose an elective pipeline: OTE, viewpoint arrangement, perspective setting location and supposition characterization. As it tends to be watched, the obstinate words are first recognized and afterward are arranged into angles. Furthermore, the stubborn piece of each perspective is delimited before playing out the notion examination. This paper is concentrated on the viewpoint arrangement and angle setting location stages and proposes a twofold commitment. To start with, we propose a half breed model comprising of a word embedding model utilized in combination with semantic closeness gauges so as to create a perspective

classifier module. Second, we expand the setting location calculation.

Online stubborn surveys are a significant wellspring of client input that organizations can use so as to gauge fulfillment and indeed, even improve their items and administrations. Additionally, client created content in sites and informal communities has tested a significant development. This has contributed to a great extent to the improvement of the Sentiment Analysis (SA) field. More solidly, Aspect Based Sentiment Analysis (ABSA) is the issue of mining sentiments from content about explicit elements furthermore, their related angles.

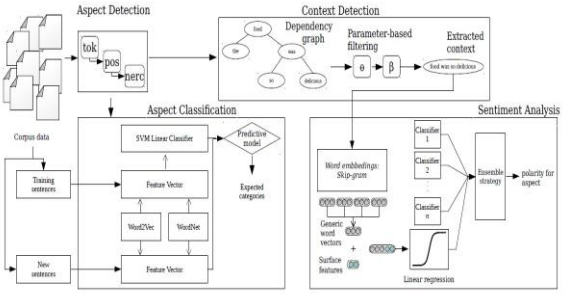


Fig. 2: Depicts the generalized flow chart explaining how the aspect-based sentiment analysis system works. [1]

The aspect category classification uses a Simple Vector Machine with a linear kernel. The output data can be printed and shown in various forms from a simple basic text based output to making of graphical representations like pie charts, flowcharts, graphs etc. context detection can be done on the dataset by the use of proper sentiment based algorithms to distinguish the dataset into various groups or categories.

The levels of sentiment analysis applied here were:

Document Level Analysis: This level groups that whether the total archive gives a positive supposition or negative feeling. The report is on single theme is considered. In this manner writings which include near learning can't be considered under archive level.

Sentence Level Analysis: The errand of this level is sentence by sentence and chooses if each sentence speaks to supposition into negative, positive, or nonpartisan. Unbiased, if sentence doesn't give any supposition implies it is unbiased. Sentence level examination is connected to subjectivity characterization. That communicates accurate data from sentences that gives abstract perspective and sentiments. For example great terrible terms.

Entity/Aspect level analysis: Both the record and the sentence level investigation don't discover people groups like and aversions. Substance/Aspect level gives all through examination. Substance/Aspect level was prior called highlight level. The centre assignment of element level is to ID builds, viewpoint level clearly gives consideration at the supposition or opinion. It depends on the idea that a feeling lives of a frame of mind and a goal of feeling.

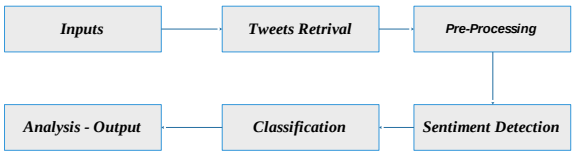


Fig.3 Depicts tweets consisting of a particular key word are collected by the algorithm employed.

After collection, the data is the freed of any inconsistency or unwanted data so that the sentiment analysis can be performed. After Sentiment estimation, classification algorithms are run on the data, supervised or unsupervised algorithm are applied to the data, depending upon the requirement. After the classification, the data we finally receive is the meaningful data which is displayed on the form of graphs or pie charts.

Another survey dwelled about how the positive reviews and the negative reviews can be used to define about a particular place or product. Many levels can be defined about a particular place regarding various parameters and scoring can be done for the place on that basis. This way the positive reviews are simplified to grade a place or product or service as to what is the most liked attribute about it, and the negative reviews are then segregated to find out the diversity in them to find out what all was disliked about the place or product. It is difficult for clients to pick products and ventures among a huge assortment of up-and-comers. An item or administration supplier has a comparable trouble in imparting its benefits to clients. With the advancement of the Web, clients can distribute their impressions and suppositions as surveys. These surveys are helpful data assets for clients and suppliers also. Specifically, the assessments of different clients, which can be affirmed ahead of time, are important for items that can't be checked by hand or administrations that can possibly be experienced if not visited there. Lodging survey is a ordinary case. An investigation of feeling examination that precisely plays out a humanly assessment regarding a matter as opposed to simply a hunt started with. In the good 'ol days, there were numerous investigations of extremity, i.e., the measurement of whether a word speaks to a positive feeling or a negative feeling. Positive/Negative table and polar expression lexicon have been developed as

the assortment of case models demonstrating feeling advances. In numerous notoriety data locales, two sorts of data are sans given portrayal and positioned assessment as for a few angles. Thusly, basic assessment investigation of whether the notoriety is sure or negative has gotten inadequate. It is important to break down which perspective is assessed positive and what is the purpose behind that. Indeed, even a similar word may have various polarities for various viewpoints. For instance, figured an assessment articulation as an example that contains descriptors. Likewise, of extremity assessment for broadly useful, it is required to investigate audits as indicated by the motivation behind use, to acquire the dependable notoriety data required by little inns. As such, applying fixed assessment criteria for feeling investigation has gotten deficient. Moreover, irregularity in the measure of positive notoriety and that of negative notoriety is perceived as a major issue on the grounds that arrangement dependent on imbalanced information as a rule makes the expectation precision lower. Accordingly, in this paper, we anticipate the extremity of every notoriety by AI what's more, assess its expectation execution, yet at the appropriate time, we center more around include word examination to realize what words portrays positive and negative audits, hoping to realize more in insight concerning how clients appraised inns. It is difficult for clients to pick products and ventures among a huge assortment of up-and-comers. An item or administration supplier has a comparable trouble in imparting its benefits to clients. With the advancement of the Web, clients can distribute their impressions and suppositions as surveys. These surveys are helpful data assets for clients and suppliers also. Specifically, the assessments of different clients, which can be affirmed ahead of time, are important for items that can't be checked by hand or administrations that can possibly be experienced if not visited there. Lodging survey is a ordinary case. [12-15]

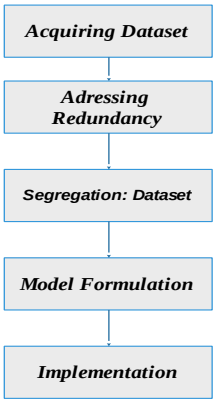


Fig.4: Implementation Phases.

Support vector machine with a linear kernel is used to perform the sentiment analysis using all the keywords in the reviews as features. Out of them the positive reviews are separated and then simplified about using the features on support vector machines with linear kernel. This classification result is then ranked and the data is published. The support vector machines have been very useful in the formulation of sentiment based models for small scale evaluations and formulations to generate the sentiment based outputs for the designated models.

Another exploration recommends the prioritization of information gathered from the inhabitants of a savvy city to clear a path for better living. The issues are distinguished from the rundown of different audits and afterward isolated in like manner. Government should have the alternative to convey and interface better for offering e-citizen upheld associations suitably to their netizens. Right now suggests passing on information through site pages and collaborate insinuates empowering occupants to give analysis and proposals. Working up a model for perception and sorting out urban organization issues from these customers made substance is a critical research issue. This paper proposes a four-advance game plan approach to manage sorting out urban organization issues from customer made substance using Aspect Based Sentiment Analysis (ABSA). Beginning, a novel jargon-based methodology is used for picking trigrams as edge articulation and pruning the overview using cross-territory stop words. Second, the resolute sentences are removed using the recognized viewpoints. Third, limit and subjectivity of the points and looking at comments close by the pack of-words for the perspectives are considered as the features for learning the characterization of the urban issue. Fourth, the thought score of a specific grouping is used to sort out the urban issues. The proposed approach is applied to a real customer created content isolated from a trade assembling on wise urban networks from an e-organization passage in India. Our investigation shows that Aspect Based Sentiment Analysis (ABSA) with principal alterations can be utilized to expel issues, proposition, and considerations from freely upheld occupant made substance (Social Data).

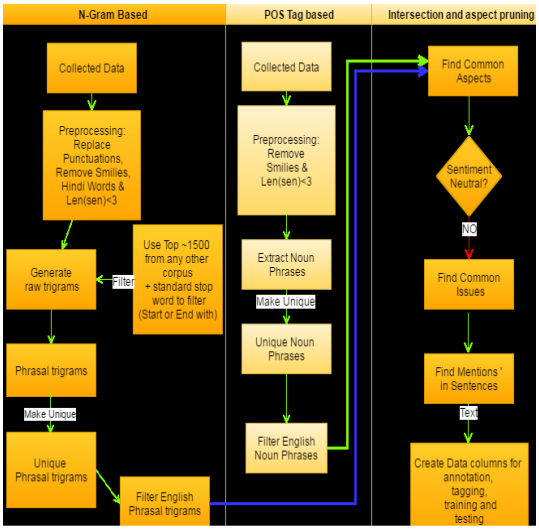


Fig. 5: Presents the schematic flowchart as to how the data flows through the model. [17]

There are various possibilities of human errors like the most common being spelling mistakes. Despite the fact that there are automated spelling checkers and correctors installed on the keyboards of mobile phones but still there are some inevitable human errors which are bound to happen. Proposed model can however overcome by proper further processing of the reviews and data from the dataset so that in the evaluation of the final output there are no or close to none cases of discrepancy coming forward in the final output for the provided dataset. [6-7, 8-11]

IV. IMPLEMENTATION

Implementation of proposed model can be summarized into following steps:

a. Acquiring Dataset

Dataset for the work was acquired from Kaggle a well-known online repository. Size of the dataset also varies according to the utility of the user who needs to apply his/her algorithm. The creation of a unique dataset is also possible but the data scientist has to take a lot of data records into consideration which can consume a lot of time and a large amount of data and processing power of the computer systems. These can be surpassed by the already created and already existing datasets, therefore a a company feedback dataset was picked from Kaggle. The tailor-made datasets are most useful in providing the best model for a specific purpose. In this project the to be used was downloaded from one of the repositories. The dataset employed for the work contained about 4 million product reviews from Amazon. This appeared to be the perfect dataset for the project because of the richness of the dataset and the variety of reviews in such a large amount.

b. Dataset Preparation: Addressing Redundancy

Redundancy in a dataset refers to the delicacy of data existing inside a dataset. The data apart from being duplicate, redundancy in a dataset can also be existent when there are pieces of data in combination of two or more pieces of data. The redundancy in the dataset is mostly not beneficial for the model can give false positive results. The application works well only when the model has been properly trained. This proper training can only be done when the dataset trains it properly and also to help it work in the most efficient manner and a non-redundant dataset is considered the most optimum for training Mostly if the data is downloaded from a repository then one can observe that the redundancy from the dataset is already checked and removed. Still, it is advised always that if it is not mentioned by the repository about any occurrence of the redundant data being checked and removed from the dataset, the data scientist should check for the redundancy and if present remove it from the dataset before using it. The work proposed employed python to address redundancy in the data acquired from Kaggle, it contained about 4 million product reviews from Amazon.

c. Segregation for processing: Dataset

Dataset obtained from the observation and recordings need to be broken into parts. This way the dataset is segregated into two or sometimes even three parts. These parts are namely : testing data, validation dataset and the training data. Among these three the testing data and training data is always created but in some cases for a more minute training of the model, validation data from the main dataset is created. The model proposed will have three informational indexes: a preparation set, an approval set and a test set. In the consequent stage, the endorsement helps with overseeing issues like overfitting, where the program may not be balanced well to manage future data. To the extent the confounding conditions that come about as a result of getting ready and test emphases. In the third stage, the test orchestrate, new test data is procured to check whether the machine continues likewise and correctly on test data as it did on getting ready data, or whether a wide delta between execution on the two stages shows overfitting has occurred. The dataset obtained from the repository was already segregated into two files for testing and training dataset due to which it was directly used in the source code of the model.

The common ratio of the dividing the dataset into the training, validation and testing dataset is generally approximately into a ratio as 30% is designated for the training dataset. The rest part is designated for the testing and the validation set. Since the validation is in principle considered to be a part of the testing dataset so we can say that it constitutes in the 70% for the dataset in working for the model.

In the model studied, it is not required to make the model obtain any local maxima, local minima, global maxima or global minima with redundancy removed there are no conditions of overfitting or underfitting, so the dataset does not need the segregation for the creation of a validation dataset, therefore Sentiment Analysis for the dataset selected was only divided into training and testing datasets.

d. Model Formulation

Computations are oftentimes made as limits; these limits fill in as meager undertakings that can be referenced by the Model proposed. By employing best fit capable computations, planners can ensure their undertakings run as fast as system support. Therefore, designs as often as possible improve existing estimations and recall them for future programming revives. The calculation will consistently be giving an appropriate assigned yield for the information embedded in dataset effectively. The calculation contemplates all the special cases in the model and how to deal with them proficiently by taking all possible variations.

The algorithm for the model is as follows:

a. Import the designated libraries for the manipulation of the data when working in the model. Here, in the model, numpy, pandas, keras, tensorflow platform, os, re, tqdm, sickit learn, bz2, pickle, etc.
b. Import the dataset from the designated folder using its path. Also add the performance measure metrics.
c. Add constants and directories.
d. Decode the utf-8 encodings, read the whole dataset, that is both the testing and training datasets. The url if present in the reviews are defined as to how they have to be manipulated.
e. The ratio of the test and training datasets is also calculated to check for the proper ratio being maintained between the testing and training data.
f. Normalize the data, maximum number of features, maximum length of each review and also the size acceptable is defined.
g. The reviews are then shuffled to train the model better.
h. Tokens are made after learning from the training dataset and these tokens are then used to segregate between various reviews in the testing dataset into positive and negative reviews.
i. After this Conv1D, BatchNormalization, padding and GlobalAveragePooling are used.
j. After these 5 epochs with a batch size of 2048 are made to fit the model after the model has been trained and tested.
k. At last, the summary and percentage of accuracy is evaluated to review the model.

V. RESULT

The library NumPy is the chief pack for consistent figuring with Python. It contains various features including these noteworthy ones:

1. A powerful N-dimensional array object
2. Sophisticated (broadcasting) functions
3. Tools for integrating C/C++ and Fortran code
4. Useful linear algebra, Fourier transform, and random number capabilities

Other than its undeniable logical uses, NumPy can likewise be utilized as a proficient multi-dimensional holder of nonexclusive information. The library Pandas is basically used for data processing and CSV file I/O. Pandas is an open source, allowed to utilize as freeware and it was initially composed by Wes McKinney. Operating with Pandas is such a less complex work with interestingly tools for records. The module OS in python traces the path mentioned on the single inverted commas to the location and then the tool shows all the objects present inside the directory or the folder mentioned as depicted in code below.

```
Out[1]: 'test.ft.txt.bz2'
```

Fig.6: Depicts the output of the commands of os.listdir.

Output shows that in the designated directory or folder the test file is present.

```
In [2]: import re
        from tqdm import tqdm
        from sklearn.utils import shuffle
        from tqdm import tqdm
        import bz2
        from keras.layers import *
        from keras.models import Model
        from keras.preprocessing.text import Tokenizer
        from keras.preprocessing.sequence import pad_sequences

        Using TensorFlow backend.
```

Fig.7: Depicts import of various modules and libraries required for the proposed model.

Most of these modules and libraries specially keras, needs a backend support of TensorFlow for the proper run and processing of the keras modules.

A standard articulation is a phenomenal game plan of characters that causes you match or find various

strings or sets of strings, using a specific etymological structure held in a model. Typical enunciations are commonly used in UNIX world. The Python module re offers full assistance for Perl-like normal enunciations in Python. The re module raises the uncommon case re.error if a bungle occurs while orchestrating or using a standard explanation. We would cover two huge limits, which would be used to manage typical explanations. To avoid any chaos while overseeing standard explanations, we would use raw strings as feedbacks or reviews.

Keras is a notable Machine Learning library for Python. It is a noteworthy level neural frameworks API fit for running on TensorFlow, CNTK, or Theano. It can run faultlessly on both CPU and GPU. Keras is a moderate Python library for significant finding that can run on Theano or TensorFlow. It was made to make executing significant learning models as speedy and straightforward as serviceable for imaginative work.

TensorFlow is a well known open-source library for world class numerical estimation made by the Google Brain bunch in Google. It can get ready and run extensive neural frameworks that can be used to develop a couple of AI applications. The bzip2 is an open-source calculation for pressure and decompression of documents. Scikit-learn is an open-source Python library that executes a scope of AI, pre-handling, cross-approval and perception calculations utilizing a brought together interface.

```
In [3]: from sklearn.metrics import precision_recall_fscore_support,accuracy_score

In [4]: def splitReviewsLabels(lines):
            reviews = []
            labels = []
            for review in tqdm(lines):
                rev = reviewToX(review)
                label = reviewToY(review)
                reviews.append(rev[:512])
                labels.append(label)
            return reviews, labels

In [5]: AMAZON_REVIEW_DIR = 'C:/Users/Lenovo/Downloads/amazonreviews'

In [6]: def reviewToY(review):
            return [1,0] if review.split(' ')[0] == '__label__1' else [0,1]

In [7]: def reviewToX(review):
            review = review.split(' ', 1)[1][:-1].lower()
            review = re.sub('\d','0',review)
            if 'www.' in review or 'http:' in review or 'https:' in review or '.com' in review:
                review = re.sub(r"([^ ]+(?<=\.[a-z]{3}))", "<url>", review)
            return review
```

Fig.8: Depicts the basic tools that are to be inserted into the program including some performance and precision metrics.

All the reviews acquired here are converted to lower case so that the same word if existing in different cases are not considered as different entities. All the url if present in any of the review are there are deleted from the review and only the review is left behind. And all the reviews and labels are defined.

```
In [8]: train_file = bz2.BZ2File(os.path.join(AMAZON_REVIEW_DIR, 'train.ft.txt.bz2'))
        test_file = bz2.BZ2File(os.path.join(AMAZON_REVIEW_DIR,'test.ft.txt.bz2'))

In [9]: train_lines = train_file.readlines()
        test_lines = test_file.readlines()

In [10]: train_lines = [x.decode('utf-8') for x in train_lines]
         test_lines = [x.decode('utf-8') for x in test_lines]

In [11]: reviews_train, y_train = splitReviewsLabels(train_lines)
         reviews_test, y_test = splitReviewsLabels(test_lines)

         100%|████████| 3600000/3600000 [01:14<00:00, 48077.53it/s]
         100%|████████| 400000/400000 [00:07<00:00, 50125.81it/s]
```

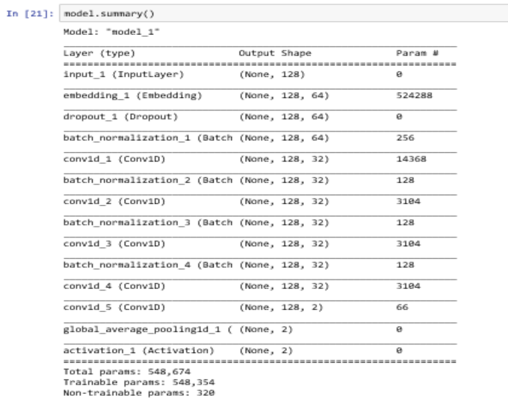
Fig.9: Bigger folder AMAZON_REVIEW_DIR is created that is made by unzipping test and training dataset files.

The utf-8 encoded strings are decoded here in both test and train files. Then these reviews are split here according to the algorithm defined. The gauge shown here is a result of the tqdm that was imported.

```
In [12]: reviews_train, y_train = shuffle(reviews_train, y_train)
         reviews_test, y_test = shuffle(reviews_test, y_test)

In [13]: y_train = np.array(y_train)
         y_test = np.array(y_test)

In [14]: max_features = 8192
         maxlen = 128
         embed_size = 64

In [15]: tokenizer = Tokenizer(num_words=max_features)

In [16]: tokenizer.fit_on_texts(reviews_train)
```

Fig.10: Code shuffles the data from the dataset in random order.

The data is stored in an array format. In Python tokenization essentially alludes to separating a bigger assemblage of content into small lines, words or in any event, making words for a non-English language. Later during the processing these tokens will be used by the model to do the sentiment analysis. The maximum length of the review is set to 128 bits. Python pickle module is used for serializing and de-serializing a Python object structure. What pickle does is that it "serializes" the article first before forming it to report. Pickling is a way to deal with change over a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information essential to revamp the thing in another python content. Pad_sequences takes into consideration the list of lists where each element is a sequence.



```
In [21]: model.summary()
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 128)               0
_________________________________________________________________
embedding_1 (Embedding)      (None, 128, 64)           524288
_________________________________________________________________
dropout_1 (Dropout)          (None, 128, 64)           0
_________________________________________________________________
batch_normalization_1 (Batch (None, 128, 64)           256
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 128, 32)           14368
_________________________________________________________________
batch_normalization_2 (Batch (None, 128, 32)           128
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 128, 32)           3104
_________________________________________________________________
batch_normalization_3 (Batch (None, 128, 32)           128
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 128, 32)           3104
_________________________________________________________________
batch_normalization_4 (Batch (None, 128, 32)           128
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 128, 32)           3104
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 128, 2)            66
_________________________________________________________________
global_average_pooling1d_1 ( (None, 2)                 0
_________________________________________________________________
activation_1 (Activation)    (None, 2)                 0
=================================================================
Total params: 548,674
Trainable params: 548,354
Non-trainable params: 320
```

Fig.11: Summary output of the processing that was done on datasets.

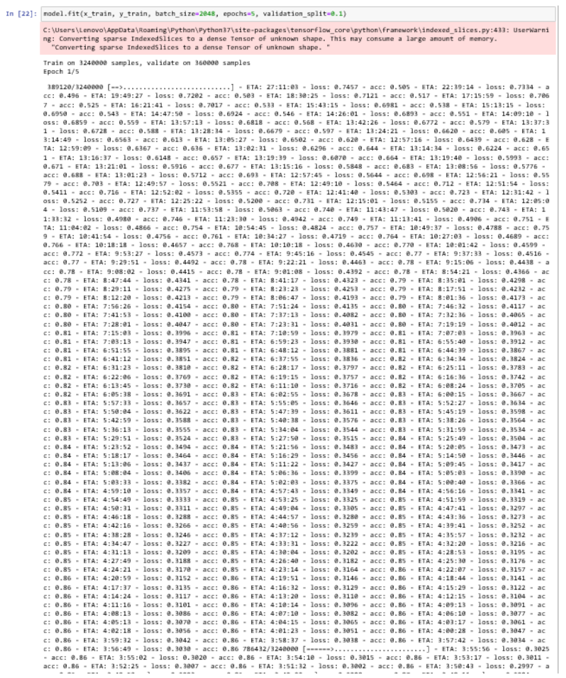


Fig.12: Result of Tokenization on the dataset

ETA is the estimated time of arrival, approximate access and loss is also mentioned. The whole data was processed 5 times in 5 epochs and with 2048 reviews together at a time. Python has characterized a module, "time" which permits us to deal with different activities in regards to time, its changes and portrayals, which discover its utilization in different applications throughout everyday life. The batch size in the epoch is taken to be 2048 because it can be evidently seen from the massive size of the dataset that 2048 reviews together will work as a thorough checker for the dataset. Any smaller batch size tends to be computationally very expensive for the system, and this will take much more time than it is taking now. It was also observed that the run time of every batch in every epoch decreases with every batch. This shows that there is a flow of the batches as all the batches are already lined up in the queue so due to this it becomes easier to access them and parse them.

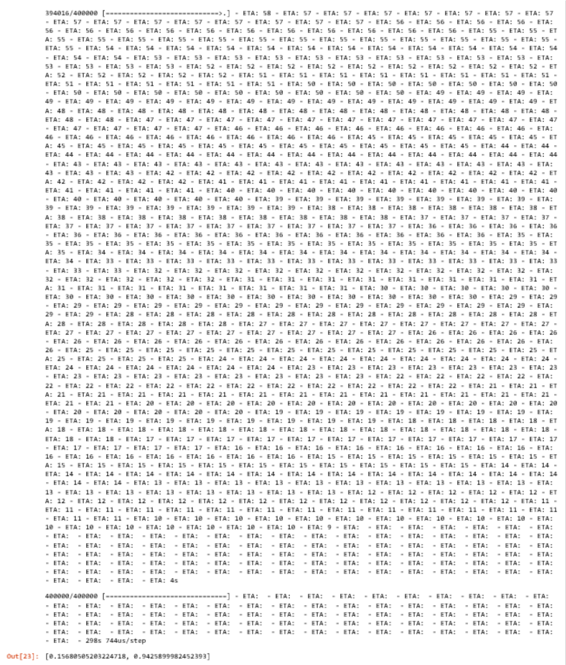


Fig.12: Diagram depicts the final output of the processing where the accuracy of the model was estimated.

Proposed model assigned tokens and performed prioritization of customer reviews via weight assignment employing sentiment analysis technique for a Amazon database obtained from Kaggle. The whole database had 4 million entries; database was subjected to cleaning process to remove redundancy. The whole data was processed 5 times in 5 epochs and with 2048 reviews together at a time. The model achieved an accuracy of 94.25% which is at par with other similar existing algorithms.

VI. CONCLUSION

With integration of disruptive technologies, i.e. AI, IoT, Space colonization, 3D printing, Blockchain, Robotics, Autonomous vehicles, and Virtual reality, sectors such as retail, marketing, corporate, finance, product development and service based all have become vastly influenced by online support systems they offer, reachability and online ratings. Customer feedback has become utmost important for the potential growth. Reviews and feedback help a lot in understanding market trends to attract customers. Owing to volume and diversity of reviews for a product - managing, segregation and prioritization of the feedbacks manually is a very tedious task especially when there are reviews and

feedbacks occurring in millions. Proposed model assigned tokens and performed prioritization of customer reviews via weight assignment employing sentiment analysis technique for a Amazon database obtained from Kaggle. The whole database had 4 million entries; database was subjected to cleaning process to remove redundancy. The whole data was processed 5 times in 5 epochs and with 2048 reviews together at a time. The model achieved an accuracy of 94.25% which is at par with other similar existing algorithms.

REFERENCES

[1] Naren.J, & Mohan, Aishwarya & Manisha.R, & Vijayaa.B,. (2014). An Approach to Perform Aspect level Sentiment Analysis on Customer Reviews using Sentiscore Algorithm and Priority Based Classification. International Journal of Computer Science and Information Technology. Vol. 5 (3). 4145-4148.

[2] Despoina Antonakaki, Paraskevi Fragopoulou, Sotiris Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks", Expert Systems with Applications, Volume 164, 2021, 114006, ISSN 0957-4174,M

[3] Abel, F., Gao, Q., Houben, G.-J., " Semantic enrichment of twitter posts for user profile construction on the social web". 2011, In Extended semantic web conference, Springer, pp. 375–389, Springer.

[4] G. Antoniou, M. Grobelnik et al, "The semantic web: Research and applications". 8th extended semantic web conference, ESWC 2011, Heraklion, Crete, Greece, May 29 – June 2, 2011, Proceedings, Part II, pp. 375–389.

[5] Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., & Gomes, J. B., "A rule dynamics approach to event detection in twitter with its application to sports and politics". Expert Systems with Applications, Issue 55, pp. 351–360.

[6] Alotaibi, F. S., & Gupta, V., "A cognitive inspired unsupervised language-independent text stemmer for information retrieval", Cognitive Systems Research, 2018, Issue. 52, 291–300.

[7] Araque, O., Corcuera-Platas, et al "Enhancing in-depth learning sentiment analysis with ensemble techniques in social applications", Expert Systems with Applications, 2017, Issue. 77, 236–246.

[8] Bhuvana, N., & Aram, I. A., "Facebook and Whatsapp as disaster management tools during the Chennai (India) floods of 2015", International Journal of Disaster Risk Reduction, 2019, 101135.

[9] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown, "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009),2009, pages 24–32.

[10] March. Luciano Barbosa and Junlan Feng," Robust sentiment detection on twitter from biased and noisy data", Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pages 36–44.

[11] Adam Bermingham and Alan Smeaton," Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage", ACM, 2010, pages 1833–1836.

[12] Michael Gamon," Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis", Proceedings of the 20th international conference on Computational Linguistics, 2010.

[13] Hamidi, H., & Moradi, S.," Analysis of consideration of security parameters by vendors on trust and customer satisfaction in e-commerce", Journal of Global Information Management (JGIM), 2017, 25(4), 32-45

[14] Conor Gallagher, E. F. , "The Application of Sentiment Analysis and Text Analytics to Customer Experience Reviews to Understand What Customers Are Really Saying", International Journal of Data Warehousing and Mining (IJDWM), 2019 , 27.16.

[15] Escobar-Rodríguez, T., & Bonsón-Fernández, R., " Analysing online purchase intention in Spain: fashion e-commerce", Information Systems and e-Business Management, 2017, 15(3), 599-622

[16] Sukesh Marla, "Sentiment Analysis of Customer Feedback on Apparel and Cosmetics Purchase in E-Commerce: A Take on Customer Satisfaction" ISSN (Print): 2204-0595ISSN (Online): 2203-1731IT in Industry, Vol. 9, No.1, 2021, pp. 76-794.

[17] Gobinda G. Chowdhury," Natural language processing", Wiley, January 2005, https://doi.org/10.1002/aris.1440370103