**PROJECT TITLE:**

**Sentiment Analysis on Tweets about lockdown using NLP**

**ORGANISATION / DEPARTMENT NAME & ADDRESS:**

Symbiosis Statistical Institute (SSI) / A 106, ICC Trade Tower 403A, S.B. Road, Pune-411016

**SUBMITTED BY:**

Student’s full name: Anamika Bhowmick, Rupali Borkar

MSc in Applied Statistics

PRN’s: 19060641006, 19060641029





**SYMBIOSIS STATISTICAL INSTITUTE**

**ACADEMIC YEAR 2019-20**

Duration of project: July-Dec,20

**Under the guidance of**

Name of the Project Guide: Akash Bhandari

Designation: GIS Engineer

Email id: akashraj54@gmail.com

Mobile: 7762931871

**Abstract:**

Individuals show feelings for ordinary correspondence. Feelings are distinguished by outward appearances, conduct, composing, talking, motions and actual activities. Emotions play an important role between two individuals. Detecting emotions through texts is a challenge for research enthusiasts. Emotion detection from the text is useful for certifiable application. The objective of this work depicted in the task is to distinguish & dissect assessment and emotion communicated by individuals from the hashtag used by them in their twitter posts about what they feel about the lockdown that happened is as yet going on and use them for investigating their opinions. We gathered tweets on one explicit point i.e lockdown and utilized the hashtag and date to extricate tweets from the necessary date and made a dataset with ID, Text, Username and Location. We utilized the dataset to identify the sentiment and emotion from tweets and on various user and tweet-based boundaries. The technique we have utilized in the paper incorporates some intriguing realities, for example, (i) Including tweets in the dataset and location, (ii) Introducing the sentiment score of tweets in influence. Emotion Detection in text documents is a content-based classification problem which involves concepts from the domains of NLP and Machine Learning.

1. **Introduction:**

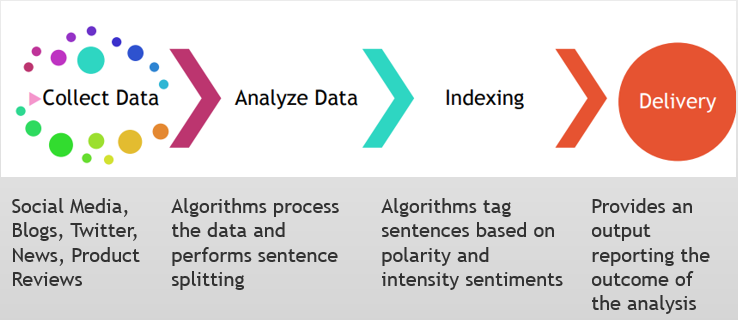
Identifying enthusiasm of an individual by investigating a text document composed by him/her seem challenging yet because major textual expressions are immediate as it consists of feeling words which results in understanding the importance of ideas of communication which are portrayed in the content archive. Perceiving the emotion content assumes existence of critical function in human-computer association. Feelings might be communicated by an individual's discourse, face expressions and written text known as speech, facial and text-based feeling separately. Adequate measure of work has been finished with respect to discourse and facial feeling acknowledgment however text-based feeling acknowledgment framework actually needs fascination of specialists. In computational phonetics, the recognition of human emotions in text is getting progressively significant from a useful perspective.

Text is an especially significant wellspring of information for recognizing feeling in light of the fact that the main part of printed information going from microblogs, messages, to SMS messages on an advanced mobile phone that has gotten progressively accessible. The quick development of feeling rich textual data makes a need to computerize recognizable proof and investigation of individuals' feeling communicated in text emotions.

Feeling is communicated as satisfaction, sadness, anger, shock, hate, dread, etc. Since there isn't any standard emotion word chain of command, centre is around the connected exploration about feeling in intellectual brain science area. W. Gerrod Parrot in 2001, created a book "Emotions In Social Psychology", in which he clarified the feeling framework and officially arranged human feelings through a feeling order in six things at essential level which are Love, Euphoria, Outrage, Trouble, Dread and Shock. Certain different words likewise fall in optional and tertiary levels. Headings to improve the capacities of current strategies for text-based emotion recognition are proposed in this paper.

Twitter information is a well-known decision for text investigation errands in light of the set number of characters i.e. 280 are permitted and the worldwide utilization of Twitter to communicate assessments on various issues among individuals, all things considered, races, societies, sexes, and so on. Here, we examined a Twitter network for emotion and sentiment detection and analysis. We recognized the feelings and assumptions from tweets and their answers and outlined an inclination network subject to messages posted by the clients. From the emotion, we perceived convincing people for both good and pessimistic feelings. Researchers have been utilizing Twitter network for different estimations and examination for quite a while. Emotions and sentiments of tweets, powerful client recognition (utilizing retweets, posts, top picks, and so on), suggestion age (in light of Twitter posts) and user influence have been tested by various specialists who have applied different strategies. In this paper, we presented some new thoughts and ideas and joined them with existing ones.

**How it works:**



1. **Literature Review:**

Numerous works has been here in this area for the most recent few years. In this segment we will see a portion of the past works done by various authors.

In [1] the creators made a corpus from Twitter utilized corpus explanation for setting up a commented-on corpus. Multi-class SVM kernels was utilized for. For highlights determination Unigrams, Bigrams, Individual, pronouns, modifiers and Dependency-parsing highlights.

In [2] the creators originally brought the tweets to make dataset. At that point they get based on target broadened highlights model. They prepared four diverse administered classifiers, Naive Bayes, SVM, ANN. SVM joined with Head Part Investigation (PCA) acquires the greatest exactness.

In [3] from the outset, the creators pre-processed the preparation dataset and took comparability estimations among the information data. At that point utilizing semantic likeness all the feeling marked corpus are bunched. In the preparation stage, the creators spoke to every text as component vector and algo is applied for preparing and train an emotion classifier.

In [4] the creator demonstrated hybrid model on emotion discovery. Here, it contains dictionary catchphrase spotting, CRF based feeling location utilizing Naïve Bayes and SVM.

In [5] the creators utilized a Concealed M.Model which decides the feeling of content. They considered sentence which contained many sub thoughts & every thought is considered as occasion that may issue a progress of a state.

In [6] the creator made a programmed feeling detection framework which can recognize feelings in tweets. His methodology includes two-section, preparing and training feeling classifier model, it depends on work from [7] & in the next part he played out a two-venture classification distinguishing tweets which contains feelings & characterizing the tweets to all the more fine-grained class utilizing delicate grouping strategies.

In [8] the creators attempted to arrange remarks with respect to a particular emergency via web-based media. They utilized feeling of outrage considering the way that equivalent strategy should be applied on different feelings also. They played out a short overview gathering 1192 reactions in which individuals are mentioned to remark under a news feature utilizing web-based media. The examination drove by Maryam Hasan et al proposed another methodology for naturally characterizing instant texts of people that construe their passionate things. Showing enthusiastic things, they used the entrenched model which describes full of feeling experience with two measurements: valence & excitement. Choosing twitter messages info informational index, it gives a huge, different and uninhibitedly accessible troupe of feelings. Utilizing hash-labels as names, their philosophy trains regulated classifiers to recognize different classes of feeling on possibly enormous informational indexes with no manual exertion and utilizing weak apparatus. Exploring the utility to a few highlights with emotion location, including unigrams, emojis, invalidations & accentuations. For handling the issue of inadequate & high dimensional element vectors of messages, using a vocabulary of feelings. They thought about exactness for a few ML calculations for characterizing twitter texts. Their procedure has an exactness of 90%, as of exhibiting vigor across learning calculations. They utilized logistic regression relapse coefficients for choosing their highlights and random forest as primary classifier.

In [9] Twitter to explore the expected utilization of web-based media to distinguish burdensome issues. Park ran a few investigations to catch the burdensome mind-set of clients at twitter. They considered around 69 people to see if their burdensome states can be reflected by their own updates. This investigation is led on three significant advances: (i) reviewing clients to recognize their downturn level, (ii) gathering tweets for the clients, & (iii) contrasting downturn level of the clients and language utilization in the tweets. It was found web-based media has helpful signs to describe the downturn in people. The outcomes demonstrated that members who has melancholy showed increment in use of the words identified with negative feelings and outrage in their tweets.

In [10] Other work which analyzes burdensome issues in people. They estimated conduct ascribes

which includes social commitment, feeling, language & phonetic, personality organization, & notices for stimulant drug. At that point they utilized these conduct highlights to assemble a factual classifier which assesses danger of sorrow. Their models demonstrated an exactness of 70% in anticipating discouragement. They publicly supported information of twitter clients who have been determined having mental issues.

In [11] Another exertion for emotion investigation on the twitter information achieved by B. and his co-workers. They attempted discovering the connection between by and large open temperament and social, monetary and other significant occasions. They separated six elements of disposition (pressure, melancholy, outrage, life, weakness, disarray) utilizing an all-encompassing variant of POMS. It was found that the social, political & monetary occasions has critical & prompt impact on different components of people’s mind-set.

In [12] As of late, Golder concentrated how singular temperament differs from every hour, every day & societies by estimating positive or worse impact in the posts, utilizing dictionary LIWC. Studies are incorporated where pooled relative proportion of antagonistic impact from RCTs could be legitimately looked at, utilizing the proportion of chances proportions, with the pooled gauge for a similar unfriendly impact emerging from observational investigations. Experimental proof from this outline demonstrates that no distinction on normal in the danger gauge of unfavorable impacts of an intercession got from meta-examinations of RCTs and meta-investigations of examinations which recommends the orderly audits for unfavorable impacts ought not be confined for explicit investigation types.

At [13] For three autonomous annotators physically coded an example of tweets and discovered nine agent classifications which includes Data Sharing, Self-Advancement, Sentiments, Arbitrary Considerations, Me Presently, Inquiries to Other people, Presence Upkeep, Stories by Me, & Tales by Others. Among them, Me Now & Irregular Musings, often showed profile proprietor's feelings were the two most mainstream classifications. To quantitatively quantify burdensome indications, the creators led a report on an individual to-individual interconnected informal community [14]. They surveyed more than 12,000 individuals consistently more than 32 years for recovering outcomes. The past perceptions gives plentiful ground for utilizing information contemplating burdensome signs as well as misery, with a definitive objective of building a constant medical care framework.

In [15] AK and TM Sebastian proposed and investigated a perspective to mine the estimation within a famous ongoing microblogging organization twitter, where customers present constant responses on and feelings on "everything". Here, they explain a half breed approach using corpus based & word reference strategies deciding the direction for feeling words in tweets. To reveal the conclusion, they removed feeling words which is a mix of the descriptors with the action words & verb modifiers in tweets. The corpus-based strategy was utilized for locating the direction of descriptive words and the word reference technique for locating the direction of action words and intensifiers. The general sentiment on the tweet then was determined utilizing the straight condition that joined feeling intensifiers as well.

1. **Proposed System:**

Here, we proposed an information based methodology and ML approach to identifying feeling or disposition of tweets

This depicts the engineering of the framework which incorporates information assortment, pre-handling, labeling, information base readiness, information approval and arrangement.

The framework depends on 2 methodologies:

1. Rule Based Methodology

2. ML Approach

1. **Methodology:**

**(a**) **Data Collection:** Twitter is one of the most well-known long-range interpersonal communication plat-frames these days with 330 million months dynamic clients and 145 million every day active users on twitter. Individuals express their feelings about their everyday live, unique social/public/worldwide issues, and so forth They share their perspectives inside 140 characters of text and at times additionally share sound/video documents. Posts are called tweets and they are public. Others can like posts, remark on them, or retweet them. Individuals can follow one another or can be Individuals’ express feelings about everyday live, unique social/public/worldwide issues, and so forth They share perspectives inside 140-character limit and at times additionally share sound/video documents. Others can like the posts, remark or retweet them. Individuals can follow one another or can be companions with one another. Not at all like most other interpersonal interaction stages, Twitter permits one directional connection, which implies one client follows another client without the last client responding the correspondence. These co-operations lead to an organization of correspondence.

The dataset which we utilized for our investigations contains an assortment of tweets, remarks and their client data. For our examinations, we picked one ongoing issue to gather tweets with different feelings. Search watchword is: #lockdown.

The dataset was set up in couple of steps – (I) gathering irregular tweets on a watchword, (ii) gathering client data (client ID, location, text), (iii) gathering remarks/answers on each tweet, (iv) gathering client data of the commentors.

Text is an unstructured data. As the data is huge, the text transforms into a representation from where we can build ML models for predicting and classifying.

We came forward many difficulties while gathering information from twitter. These are:

(i)A few shared photographs & recordings and didn't specify in the content.

(ii) In any event, for tweets which communicated as text in English, loads of remarks were in other languages.

(iii) Bunches of remarks had no content, they just shared photographs or recordings.

(iv) A few remarks on certain tweets were not from an individual, but rather from certain records of information channels or business people. Those remarks were essentially notice of some news or item.

(v)Most clients didn't share their area.

(vi) Barely any individuals had a huge number of tweets, however none of them were unique.

(vii) Scarcely any answers didn't have any content other than couple of notices of certain records.

(viii) Not many answers just had the equivalent hash tagged word as the first tweet and nothing else.

(ix) Barely any tweets and answers just expressed a few realities without communicating any notions or feelings.

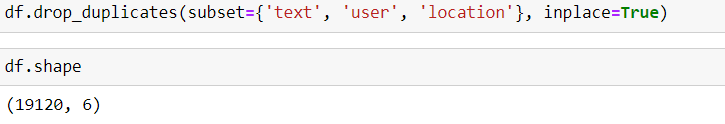
(x) Barely any answers had emoticons and no content.

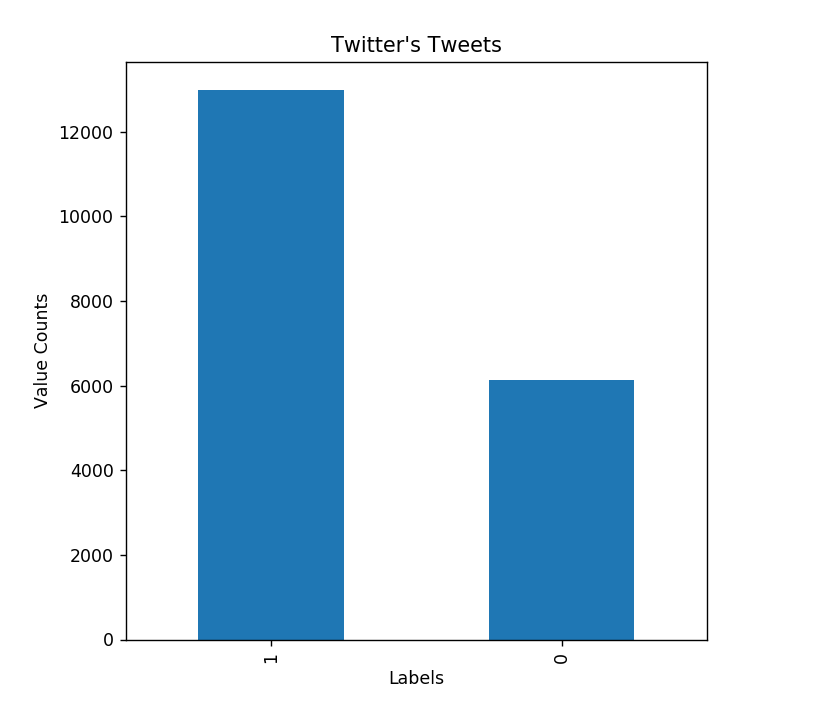
(xi) Some repliers reacted by asking questions which didn't communicate any feeling.

(xii) A few reactions were absolutely arbitrary and outside of any relevant connection to the subject at hand.

We used twitter API to scrap the data for the data collection. Because of the rate furthest reaches of twitter Programming interface i.e. API, 15 Programming interface calls are permitted in each 15 min which limited the measure of gathered information. We gathered and annotated 20000 tweets and answers from 23rd July 2020 to twentieth August 2020. As per the tweets and answers, we gathered data for 20000 clients. As we manually clarified each tweet and answer as per feeling, slant and understanding score, the dataset had restricted measure of information. We labelled the text as 1 if the tweet was positive and 0 if the tweet was negative and this was done manually so it took around 1 and half month to label 20000 tweets. In this paper, we attempted to venture out a customized informal organization recommender by proposing another methodology for Twitter feeling and notion network.

In the Basic EDA as the dataset consists of 20000 tweets and there are 4 attributes of it which are UserID, Text, Location and Labels. Started doing the analysis with cleaning dataset by using ‘data deduption’, which removes duplicates. After doing data-deduption we had 19120 data.



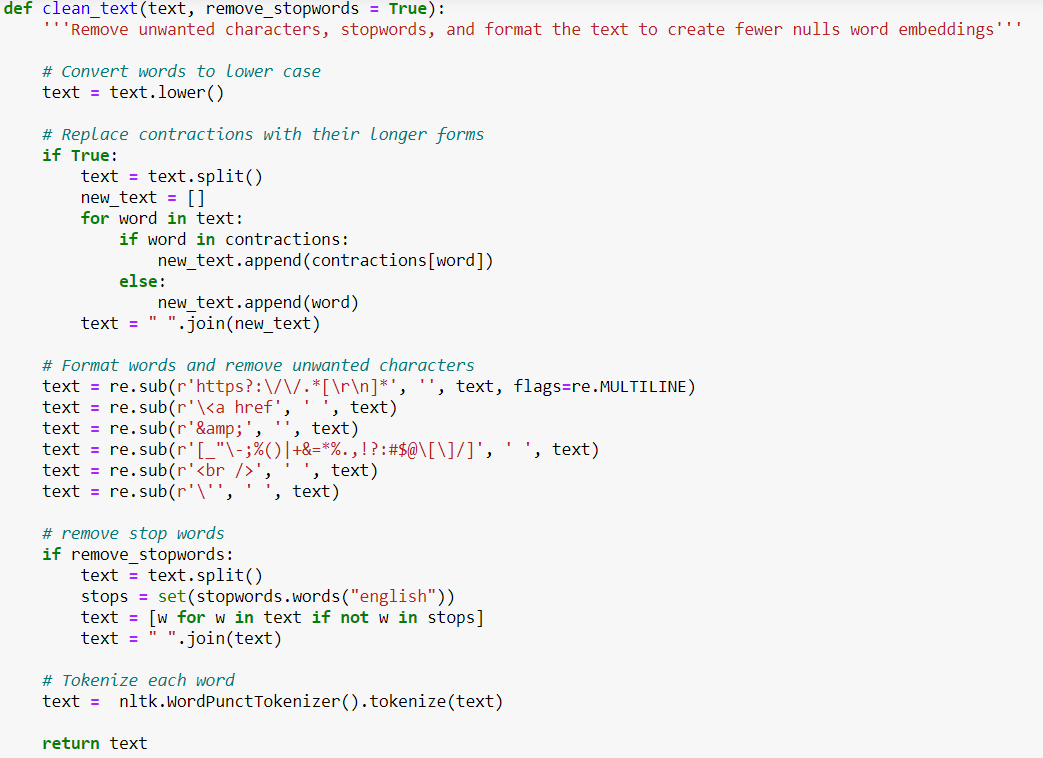


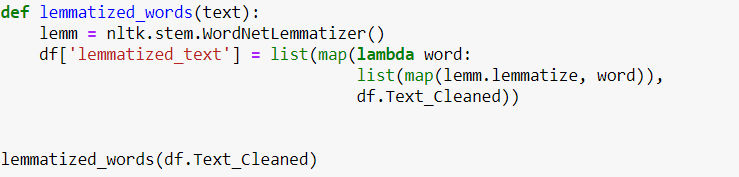
**(b)** **Text Cleaning and Data Pre-processing:**

There can be numerous methods of cleaning and pre-processing the literary information and here we have applied the ones which are much of the time utilized in NLP pipelines.

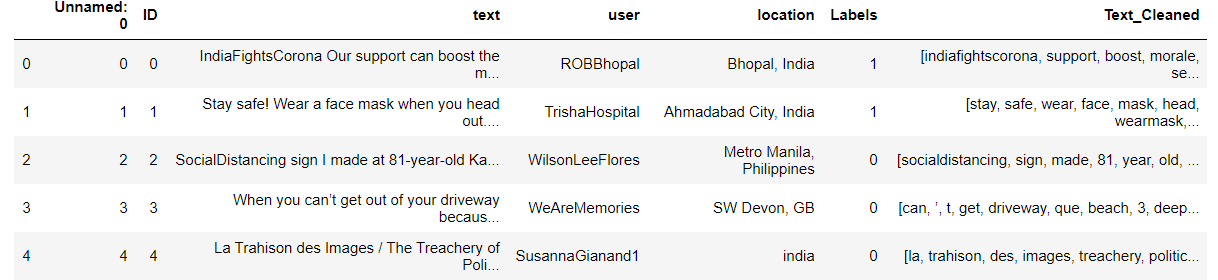
* Lower case: We hope to treat "Lockdown" & "lockdown" as a similar word, without making different anticipating powers, we've down-cased each word.
* Contractions: We have supplanted withdrawals with their more drawn-out structures, for example, "isn't": "isn't", "can't": "can't". To do as such, I've imported withdrawals list from stack over-flow.
* Eliminate extraordinary characters: We’ve cleaned the information from any unique character, for example, twofold statements, accentuation.
* Stopwords: We have eliminated stopwords as they add clamor without bringing any data esteem in demonstrating. We've downloaded rundown of English stopwords from the nltk bundle and erased them from the content corpus.
* Tokenization: To handle text, we have to part it to more modest pieces. Here, we have part sentences into words utilizing WordPunctTokenizer from the nltk library. Tokenizer partitions content to a rundown of sentences and word tokenizer which helps in isolating sentences to words. Basic tokenizer isolates the strings into substrings utilizing the string split() technique. Tokenizing utilizing a specific delimiter string that utilize the split() technique straightforwardly, as this is more productive. Straightforward Tokenizer isn't accessible as isolated capacity; all things being equal, we can simply utilize the split() strategy legitimately. Utilizing split technique with space as delimiter we have made the substance into tokens. The entire sentence is converted into tokens".
* Lemmatization: To change over each word into its root word, we've utilized Lemmatizer.

This is the text processor used to apply all the steps listed above:





Here’s a look of 5 processed datasets:



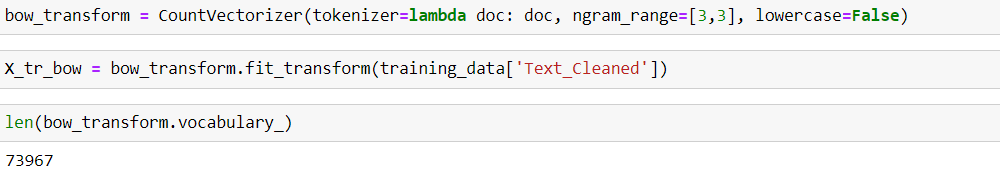
**(c)** **Feature Engineering:**

Since ML models don't acknowledge the crude content as input data, we have to change over "Tweets" into vectors of numbers.

There are various methods of changing content into numeric vectors. Here, we have applied first Bag of Words, trailed by Bag of-n-Grams, and later I've to Tf-Idf. We have intended to demonstrate two distinctive arrangement by utilizing these procedures & analyze their exhibitions on the tweets extricated.

* **Bag of Words (BoW):**

It is a basic yet extremely successful method of speaking to message. It has extraordinary accomplishment in language demonstrating & text grouping. It depends on the word check insights.

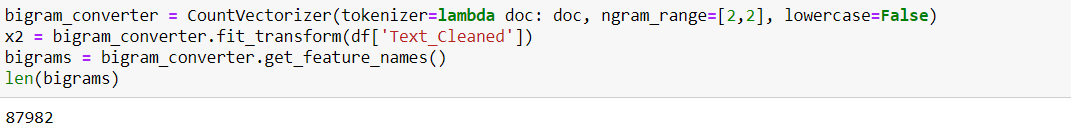


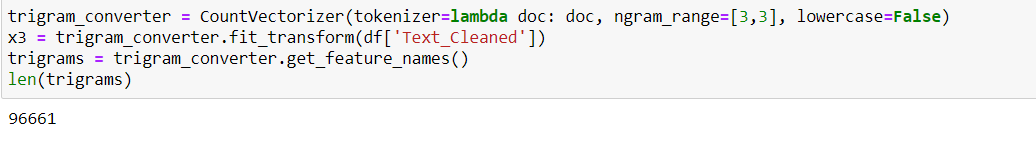
BoW(w, d)= Number of times the word w appears at document d

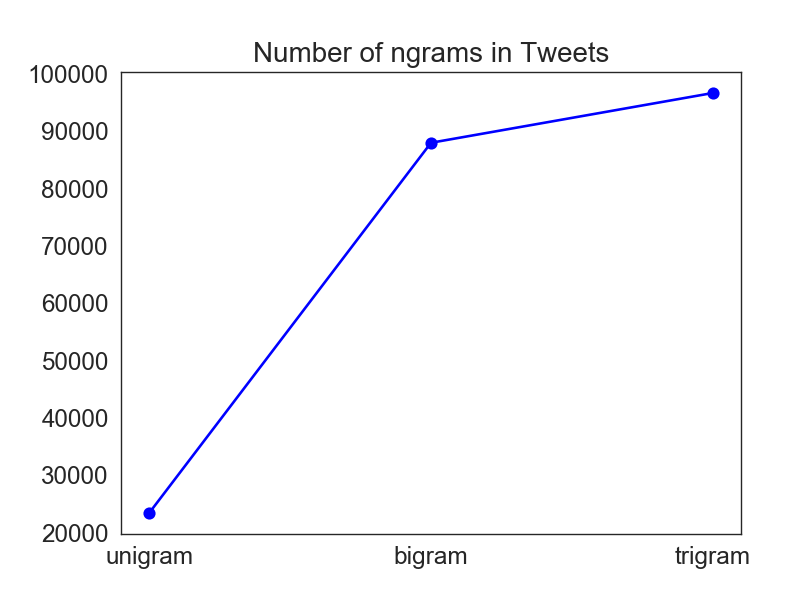
Since bag of-words portrayal changes over content into a "level vector of numbers", it doesn't recall any unique literary arrangement and it can pulverize the semantic importance of the content. As such, it just records how frequently each word shows up in the content and it doesn't give any significance to their request. Each word tally turns into a measurement for that particular word.

* **Bag of n-Grams:**

It is an augmentation of BOW & speaks to n-grams as succession of n tokens. All in all, a word is 1-gram (unigram), two words are 2-grams (bigram), and so on is applied in NLP pipeline since it holds first succession of content more than the Sack of Words portrayal. In any case, it has an exceptionally high computational expense, in light of the fact that hypothetically k remarkable words can mean k² interesting bigrams.







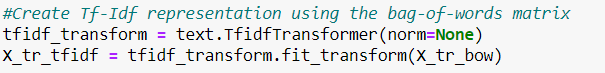
* **Tf-Idf:**

Tf-Idf represents term recurrence converse record recurrence, and as opposed to ascertaining the includes of each word in report of the dataset (Bow), it figures the standardized tally where each word check is partitioned by the quantity of archives this word shows up.

Tf-Idf(w, d)= BoW(w, d) \* log(Total Number of Records/(Number of archives in which word w shows up))

On the off chance that a word shows up regularly in a specific archive, however not in so numerous different records, all things considered, the word speaks to a specific importance for that report and gets a bigger tally gratitude to high Idf. On the opposite side, in the event that a word shows up in numerous records, at that point its Idf is near 1 and the logarithm transforms 1 into 0 and diminishes its impact.

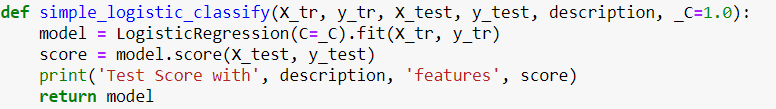
.



* **Logistic Regression:**

After making a 70/30 train-test split in data, we have applied strategic relapse which is grouping calculation used to take care of classification issues. The LM classifier utilizes the weighted blend of the information highlights and goes them through sigmoid capacity. The capacity changes any genuine number contribution, to a number somewhere in the range of 0 and 1.

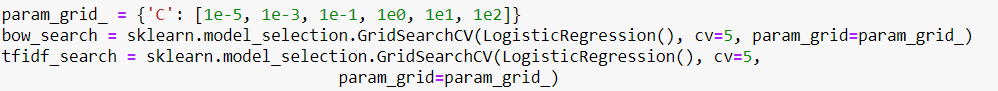




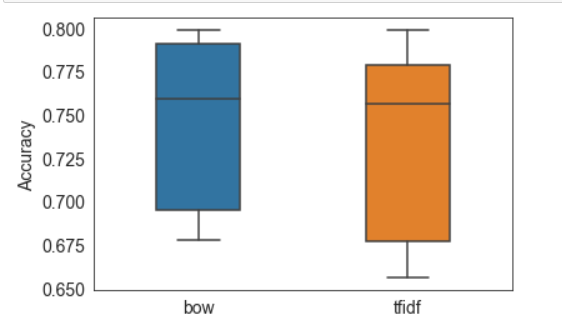
We applied LM classifier, on both Bag-of-triGrams and Tf-Idf highlights to analyze their exactness scores. Building the models on the default boundaries gives us the precision scores as underneath:



Nonetheless, when quantity of highlights is higher from quantity of information focuses, the model will in general be underdetermined. To fix this issue, we have to present extra requirements that are known as hyperparameters. By utilizing GridSearch we had the option to attempt various blends of qualities to locate the model with the most reduced blunder metric, which is for this situation log misfortune. In LM 'C' decides the measure of regularization, and the lower esteems increment regularization.



Building the LM classifier with various regularization boundaries utilizing GridSearch, we've prevailing to improve exactness scores.



As referenced previously, our ML model probably accomplished a higher score than 78% to give preferred execution over anticipating the survey consistently sure. By utilizing tuned strategic relapse classifier on both, we have acquired victories. It is additionally essential to feature that Tf-Idf featurization performed slight better contrasted with pack of-trigrams which has likewise accompanied an exceptionally high computational expense.

Despite the fact that from this precision scores, it may have appeared to be that we didn't build the exactness much… to assess this speculation, we have tested to prepare the model by similarly dispersing the positive and negative surveys. Presently we a more modest dataset however it has adjusted classes so the beginning likelihood is half!

We applied LM classifier with hyperparameters tuned by utilizing GridSearch & we had the option to acquire 79% precision which contrasted with half beginning stage is looking much better.

1. **Results:**

In this part, we will talk about our outcomes subsequent to exploring different avenues regarding our proposed model. We have taken 20000 tweets twitter by web scraping using Twitter API. The quantity of information diminishes a great deal by eliminating loud tweets during pre-preparing and separating them through WordNet feeling words. We have found 17000 people thought that the Lockdown was positive and 3000 peoples found that the Lockdown was negative. We have expected that one tweet has just a single feeling class and just a single feeling word in the tweet speak to that feeling.

After pre-preparing and separating the tweets for feeling word, we split them into preparing and testing dataset to prepare and run them into customary ML Classifiers.

1. **Conclusion:**

Emotion Discovery can be viewed as a significant field of exploration in human-PC association. An adequate measure of work has been finished by analysts to distinguish feeling from facial and sound data though perceiving feelings from literary information is as yet a new and hot exploration region.

Emotion discovery is perhaps the hardest issue to unravel. Negative feeling from text is a difficult work and most. of the examination works have some thoughtful restrictions most importantly, language uncertainty, different feeling bearing content, text which doesn't contain any feeling words and so on However we have attempted a few ways to deal with distinguish feeling from twitter. Our impediments are that we have utilized a little example as our dataset and there are still language uncertainty issues as we have not had the option to address messages which speak to different feeling simultaneously. Later on, we will acquaint Deep Learning methods with distinguish feeling identification on this dataset.

With the proposed framework we can identify and group the feeling of the tweet. The framework execution has been expanded in the second methodology that is ML approach when contrasted with the main methodology that is Rule based methodology. Here, strategies which are as of now being utilized to identify feeling from text are assessed.

1. **References:**

[1] R. C. Balabantaray, M. Mohammad, and N. Sharma, “Multi-class Twitter Emotion Classification: A New Approach,” International Journal of Applied Information Systems, vol. 4, no. 1, pp. 48–53, 2012.

[2] M. Anjaria and R. M. R. Guddeti, “Influence factor based opinion mining of Twitter data using supervised learning,” in 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp. 1–8.

[3] S. Yuan, H. Huang, and L. Wu, “Use of Word Clustering to Improve Emotion Recognition from Short Text,” Journal of Computing Science and Engineering, vol. 10, no. 4, pp. 103–110, 2016.

[4] H. Yang, A. Willis, A. D. Roeck, and B. Nuseibeh, “A Hybrid Model for Automatic Emotion Recognition in Suicide Notes,” Biomedical informatics insights, vol. 5, p. 8948, 2012.

[5] D. T. Ho and T. H. Cao, “A High-order Hidden Markov Model for Emotion Detection from Textual Data,” in Pacific Rim Knowledge Acquisition Workshop, 2012, pp. 94–105.

[6] M. Hasan, E. Rundensteiner, and E. Agu, “Automatic emotion detection in text streams by analyzing Twitter data,” International Journal of Data Science and Analytics, vol. 7, no. 1, pp. 35–51, 2019.

[7] M. Hasan and E. Rundensteiner and E. Agu, “Emotex: Detecting Emotions in Twitter Messages,” 2014.

[8] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu, May 2014, “EMOTEX: Detecting Emotions in

Twitter Messages,” ASE

BIGDATA/SOCIALCOM/CYBERSECURITY Conference, 27-31.

[9] Minsu Park, Chiyoung Cha, and Meeyoung Cha, 2012,“Depressive moods of users portrayed in twitter,” in

Proc. of the ACM SIGKDD Workshop on Healthcare Informatics, HI-KDD.

[10] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz , 2013, “Predicting depression via social media.,” in International AAAI Conference on Weblogs and Social Media (ICWSM'13), The AAAI Press

[11] Johan Bollen, Huina Mao, and Alberto Pepe, 2011,

“Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena,” in International AAAI

Conference on Weblogs and Social Media (ICWSM'11).

[12] Golder S, Loke YK, Bland M, 2011, Meta-analyses of Adverse Effects Data Derived from Randomized

[13] Munmun De Choudhury, Scott Counts, and Michael

Gamon, 2012, “Not all moods are created equal! Exploring human emotional states in social media,” in

Sixth International AAAI Conference on Weblogs and Social Media (ICWSM'12).

[14] Carlo Strapparava and Rada Mihalcea, 2008, “Learning to identify emotions in text,” in Proceedings of the 2008

ACM symposium on Applied computing. ACM, pp.1556-1560.

[15] Akshi Kumar and Teeja Mary Sebastian. 2012,

“Sentiment Analysis on Twitter”, International Journal of

Computer Science Issues, Vol. 9, Issue 4, No. 3, ISSN (Online): 1694-0814