iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** III **Month of publication:** March 2020

DOI: http://doi.org/10.22214/ijraset.2020.3066

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

Understanding Viewers Sentiment on Skin Whitening Glutathione Product Review using YouTube Comments

Mohammed Azam Sayeed[1], Noor Ayesha[2]
[1]Department of Computer sciences, SET Jain University Bangalore -562112

Abstract: Skin whitening creams do big business among crowd of Indian decent and have serious mental, physical and emotional health implications for those who use these products. Whiteness theory and the color complex leading to studies of "colorism" provide a theoretical backdrop by which to understand some of the root causes of the desire for fair skin. This paper explain the methodology used to extract and understand the overall sentiments of indian on using glutathione for skin whitening through youTube comments.

Keywords: Skin whitening, sentiment analysis, glutathione, selenium, textblob, data visualization, custom classifiers, Deep Neural networks, TF-Hub, Naïve bayes classifier

I. NEED TO UNDERSTAND THE PRESENT SENTIMENTS OF INDIANS ON SKIN WHITENING

[1]There is extensive documentation of the negative impact of skin whitening on both mentally and physically constituting a large public health issue within this population. According to a 2006 edition of Harper's Magazine, the number of new skin-whitening products that have been introduced to Asian and Pacific markets exceeds one-hundred and eighty-nine since 2002 ("Events about Race"). Skin whitening creams do big business in Hong Kong, the Philippines, Japan, India and Malaysia and yet most have been proven to cause great physical harm to those who use them (Jones, 2004). The skin whitening creams themselves lead to severe physical deformities (from ingredients such as mercury and hydroquinone which are known to burn skin) to feelings of inadequacy if one's natural skin tone is perceived as "dark" (Saifee, 2005). In addition to physical health consequences, the desire for "whiteness" has major psychosocial implications including lower self-confidence levels as well as perceptions of inferiority among Indian women (Saifee, 2005; "India Debates" 2007). The embedded desire for fair skin plays out among the Indian community in the form of a color bias, which prizes lighter skin over darker skin.In 1901 Risley commissioned an all India Census using caste as a primary category. The British East India Company used this census to argue the importance of caste which became an important method of classification. In Risley's words, "The principle of Indian caste is to be sought in the antipathy of the higher race for the lower, of the fair-skinned Aryan for the Black Dravidian." (Prashad, 2000) Fair skin color and race became a basis for caste difference and religious differences were seen as supportive of the entire system. The Census was used to determine where caste groups mobilized, to support the use of segregated land and quotas for various non-Brahmin peoples and to control electoral representation, appointments to government jobs as well as entry to educational institutions (Dirks, 2004).The legacy of colonialism perpetuated a national identity of subordinate Indians which still today expresses itself through images of "whiteness" as superior to "darkness (Chakravarty, 1989). Internalized racism within Indian society shows strong links with the notion of a color complex. Among African-American populations, the color complex refers to lighter skinned Blacks rejecting Blacks who are darker (Russell, Wilson & Hall, 1992). This complex reveals itself in a variety of situations from work environments to social situations.

Racism has become a part of the Indian psyche. The reality is that an overwhelmingly brown nation looks down on those that are dark skinned. Indian anthropologists say the preference is ancient, carved into the culture by waves of light-skinned invaders, most recently the British that left natives feeling inferior. Theorists argue that whiteness is not a biological classification but a social construction which provides material and symbolic privilege to whites and those who pass as white.

Within Indian communities, the major female actors in Bollywood (the film industry in India), for the most part, are not representative of the average Indian woman. But have extremely fair skin and European features (McPhate,2005). The implication is that fair skin and light eyes are the representatives of women in India, when the majority of the population both in India and abroad, do not share these characteristics.

A. The Media Ideal Of Beauty

In early 2006, Hindustan Lever Limited, the parent company which produces the popular skin lightening cream Fair and Lovely, produced a series of television advertisements targeting dark skinned women. One of the most notable (and thus controversial) depicts a dark skinned woman and her father. The father complains to his wife "Wish we had a son" The daughter overhears this, uses the skin lightening cream and the final sequences show her landing a good paying job as an air hostess as epitome of success. This is one of many advertisements that further perpetuate both a sexist and colorist view of the way men and women should be. Many ads influence the minds of young children in India to believe success and happiness is achieved only through the beauty standards of lighter skin tone. Research has found that the preference to lighter skin tone is so much rooted in the subconscious mind of Indians , shift in mentality will take time and patience. Whiteness is being sold as a cosmetic product, an 'effect' you can buy and put on" (2003). The advertising industry makes large amounts of money that cater directly to darker skinned women with the assumption that creams will improve their class status. One recent study of magazine advertisements in four countries, including India show that the depiction of women focuses on gender roles and beauty ideals. Hindustan Lever Limited is India's largest company delivering personal products to consumers and has the largest advertising budget at $29 million annually. It is important to note the double standard for advertising in India and in the western world as it also own dove which spread the message to embracing all skin tones as beautiful. The dichotomy in how particular brands are marketed ultimately comes down to the amount of money the multinational corporation can make (Sinha, 2000; Karnani, 2007). Economically, fairness creams are marketed to the Indian community because they create a profit of almost a quarter of a million dollars per year in India alone and gives Unilever a 53% of the fairness cream industry (Challapalli, 2002).

B. Physical, Emotional And Mental Health Aspects Of Skin Lightening

Many skin lightening creams have devastating physical health repercussions. Skin whitening products work in various ways. Some use acid to remove older, darker skin to reveal lighter skin underneath. Others inhibit melanin production; the chemical that produces the color in skin tone and contain mulberry extract, kojic acid, arbutin and hydroquinone (Fuller, 2006). Most popular bleaching creams contain 4% hydroquinone, a chemical also used in photo Processing and rubber production that can take off the entire outer layer of the skin. Creams with these chemicals have been shown to cause itching, burning and blistering also, eventually leading to darker skin patches where the creams were administered (Abaas, 2006). There is also some evidence that the long term effects of such creams may cause skin disorders from simple irritation to skin cancer (Dussault, 2006). Cosmetic creams containing hydroquinone have been banned in both the United Kingdom and the United States and the FDA hopes to ban the remainder of the skin bleaching creams (Sinha, 2000).Whitening creams pose a greater risk in countries such as India where there are no central regulations of the industry. Fair and Lovely along with other creams, soaps and lotions are often sold in corner drug stores, markets and administered by beauticians. This trend extends into the Western world as well. An informal analysis of Indian stores in California yielded a 100% availability of Fair and Lovely creams for purchase over the counter. Such creams are not considered pharmaceutical products in India or abroad and are therefore not subject to testing and regulations, making the long-term effects of these products largely unknown (Karnani, 2007). Mental and emotional health effects appear better documented. Arun Adhikari, executive director for personal products at Hindustan Lever, said "Historically Fair & Lovely's thoroughly researched advertising depicted a before and after effect". In reference to the current television advertisements, Adhikari notes, "[We] show a negative and positive situation. We are not glorifying the negative but we show how the product can lead to a transformation, with romance and a husband the pay-off" (Luce and Merchant, 2003). Their website boasts the slogan "Guaranteed fairness, guaranteed fame (HLL, 2007)." Hindustan Lever's research says 90 percent of Indian women want to use whiteners because it is "aspirational, like losing weight". A fair skin is, like education, regarded as a social and economic step up (Luce and Merchant,2003).

These types of messages, especially when aimed at young women who are already subjugated in Indian society have a marked negative impact on their emotional health. Fair and Lovely's advertising campaigns are racist and perpetuate the notion that fairness is equivalent to beauty ("India Debates", 2007). Recent research indicates that the target audience for Fair and Lovely is predominantly women between the ages of 18 and 35 years old and further evidence that school girls in the 12 to 14 year age groups widely use fairness creams (Karnani, 2007). Negative media images have a profound impact on the way young Indian women identify themselves (Dagar, 2004). Smitha Radhakrishnan reiterates the double standard between men and women which also plays a role in the emotional development of women.

The more women focus on things like beauty and attracting men, the more we are distracted from the things that are truly important. The less they can focus on things like avoiding harassment, equal opportunities, empowerment, improving self esteem and finally acceptance of oneself.

C. Glutathione Tablets For Skin Whitening

Currently there is a lot of hype and usage of glutathione of skin whitening, the problem is there is no concrete study done on its actual effects for skin lightening but easily available on the counters. There is no extensive documentation on long term side effects, being very expensive. The problem on using this is the real issue is not whether it actually help in skin lighting but worsening the deep Emotional mindset on people by fairness products. Indians have been stuck in this vicious cycle of achieving impossible beauty ideal standards for very long time and endorsing glutathione is only worsening the situation. In this paper we will understand how Indians react to glutathione product endorsement and if there is an actual shift of strong rejection of skin whitening or are people still willing to invest in this pitiful beautiful business. Study of sentiment using YouTube video on Endorsing glutathione for skin whitening makes it more meaningful at the present times as this drug is relatively new and will represent the sentiment of current youth generations.

II. PROPOSED METHODOLOGY FOR SENTIMENT ANALYSIS

A. Data Collection

We collect the comments from YouTube video which favors/markets the use of glutathione tablets for increasing fairness for Indian skin tone. The video was selected based on filter on most watch video, and title names featuring Best Glutathione Tablets in India, to understand the sentiment of the Indian viewers perceiving the drug review. The selected video has around 7,800,000+ views currently and has lot of comments from Indians expressing their views. For collecting comments from YouTube of specific video link is achieved using Selenium framework.

Selenium Framework is a suite of automation testing tools that is based on the JavaScript framework. It could run the tests directly on the target browser, drive the interactions on the required web page and rerun them without any manual input.

This eliminates repetitive manual testing that consumes lots of time and effort. Basic Test automation of websites functionality using selenium on web browsers is automated to proceeding to subsequent pages web pages or hyperlinks by automating manual clicking of buttons and filling text fields.

1) *Installation of Selenium:* Since we are using Chrome browser, we need to install Selenium framework in python environment with the corresponding suitable DriverManager to interact with the correcponding web browser without intervention of human interaction and perform the test automation.

The following commands are required for extracting comments

>> pip install –U selenium

>>pip install webdrivermanager

```
import time
from selenium import webdriver
from selenium.webdriver import Chrome
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from webdriver_manager.chrome import ChromeDriverManager

i=0
with webdriver.Chrome(ChromeDriverManager().install()) as driver:
    wait = WebDriverWait(driver,10)
    driver.get("https://www.youtube.com/watch?v=XXXXXXXXX")

    for item in range(500): #by increasing the highest range you can get more content
        wait.until(EC.visibility_of_element_located((By.TAG_NAME, "body"))).send_keys(Keys.END)
        time.sleep(3)

    for comment in wait.until(EC.presence_of_all_elements_located((By.CSS_SELECTOR, "#comment
#content-text"))):
        print(i,comment.text)
        i=i+1
```

Fig. 1 Extracting youtube comments using selenium and DriverManager

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

The above code the most important statement is
>>with webdriver.Chrome(ChromeDriverManager().install()) as driver:
Which first looks for latest chromedriver installed in system , if not found installs the correct chromeDriver .exe file in System
Output message in our case:
Looking for [chromedriver 80.0.3987.106 win32] driver in cache
File found in cache by path [C:\Users\mosa0816\.wdm\drivers\chromedriver\80.0.3987.106\win32\chromedriver.exe]
The YouTube video is specified in driver.get method as below
>> driver.get("https://www.youtube.com/watch?v=XXXXXXXXX")
To read more content ,we can increase the range limit, this for loop is responsible to locate the body and scroll the webpage correspondingly to the number of scrolls mentioned in the range limit
>> for item in range(500): #by increasing the highest range you can get more content
>>   wait.until(EC.visibility_of_element_located((By.TAG_NAME, "body"))).send_keys(Keys.END)
>>   time.sleep(3)

And the last for loop extracts the comments using css properties for sentiment analysis.

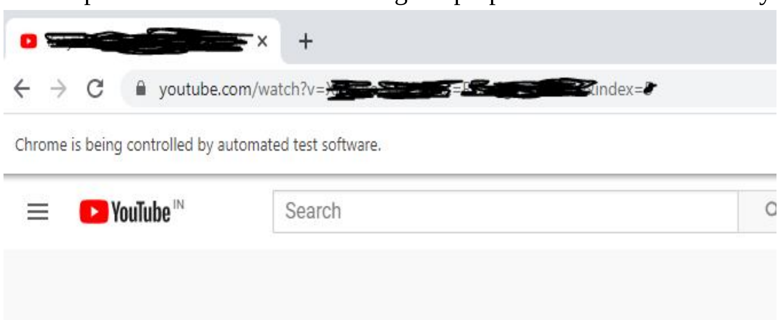


Fig. 2 Test automation by chromeDriver for automated scrolling and extraction of comments

2) *Saving Comments in Internal File:* We have extracted about 500 comments from above step and stored internally in file for further steps of analysis, here since most of the viewers are of indian origin, some of the comments are written in English but interpretable language is hindi, we remove these comments in the internal file. Also All Emoji's symbols were removed using the Regular expression replace functionality in notepad++ with RE as [^\x00-\x7F]+, also optionally common sms language such as y,ua,ur,plz etc can be replaced in notepad++, but not mandatory as will be done in next steps for data preprocessing.

B. Data Preprocessing

Most of the data preprocessing on extracted youtube comments is achieved by the TextBlob API
From documentation of API: TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. TextBlob aims to provide access to common text-processing operations through a familiar interface. You can treat TextBlob objects as if they were Python strings that learned how to do Natural Language Processing.

1) Installation for Data Preprocessing

For using Textblob features
>>pip install –U textblob
Next we need nltk package for usage which can be installed in python notebook as
>>import nltk
>>nltk.download()
Which will download all necessary modules related to nltk using GUI installation
After installation is complete, set the environment variable NLTK_DATA having path of installed nltk such as C:\nltk_data.
Lastly we also install gingerIt, It is used for Correcting spelling and grammar mistakes based on the context of complete sentences.
gingerIt is a Wrapper around the gingersoftware.com API
>>pip install gingerit

2) *Reading from Internal file and Converting Entire Comment Text to lower case with removal of punctuations:* Next step we read the entire text in Internal file for pre-processing, We can read the text from file using the file operations API provided by python , next str.lower() function converts the entire fetched text to lower case and then we remove unnecessary and redundant punctuations such as !!!,…,??? etc. Lastly we split the text into individual comments using the split() method on '\n' new line.

```
from textblob import TextBlob
from gingerit.gingerit import GingerIt

with open('comments.txt', 'r') as f:
all_text =f.read()

all_text =all_text.lower()

from string import punctuation
t=all_text
all_text_withoutPunc = ''.join([c if c not in punctuation else ' ' for c in t ])

reviews_split = all_text_withoutPunc.split('\n')
print ('Number of reviews :', len(reviews_split))
```

Fig. 3 Lower case and removal of punctuation

After splitting the text, we have got 300 individual reviews to study the sentiments of Indians on glutathione for skin whitening.

3) *Spelling Correction using gingerIt:* We can correct spelling of individual indexed review, using Textblob.correct() method or gingerIt, we chose the later as gingerIt has much more accuracy than textblob. correct(), Also after correction by gingerit it may add few punctuations like commas and Question marks based on grammer, we retain the Question mark punctuation and replace the added commas using the str's replace method

```
review = reviews_split[1]
parser = GingerIt()
#line==string you wanna correct
tweet=parser.parse(review)
                                        blob =TextBlob(tweet['result'].lower())
                                                 blob=blob.replace(',','')
```

Fig. 4 Spelling correction using gingerit

4) *Translating And Hindi Language Detection And Removing Spaces In Beginning And End Of Review:* As Explained earlier we are understanding the sentiments of Indian on use of glutathione for skin whitening, there is a possibility of the comments to be written in Regional spoken language as Hindi. Hence we use textblob for Hindi language detection and correspondingly translating to English language , translation is quite accurate and retains the message intended to convey with appropriate spelling.

```
if blob.detect_language() =='hi':
   blob.translate(from_lang="hi", to='en')

blob = blob.strip() # or use lstrip() and rstrip()
```

Fig. 5 Removing spading in comments

C. Sentiment Analysis using TextBlob: The sentiment property returns a named tuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. We collected the Polarity and subjectivity value and corresponding classified class (pos/neg) using the Textblob and stored in the .CSV file for further steps of data comprehension. This is done for each data pre-processed comments iteratively using below code snippet.

```
In [84]:
for i in range(300):

    review = reviews_split[i]
    tweet=parser.parse(review)
    blob =TextBlob(tweet['result'].lower())
    blob=blob.replace(',','')
    blob = blob.strip()
    #print('review at ',i)
    if blob.sentiment.polarity >0:
        class1='pos'
    else:
        class1='neg'
    print(blob,',',round(blob.sentiment.polarity,2),',',round(blob.sentiment.subjectivity,2),',',class1 )
```

```
clear skin is more important than lighter skin  clears your skin that's enough , 0.25 , 0.6 , pos
i'm proud of my dark skin i don't want to be fair you should be happy with what you are , 0.54 , 0.82 , pos
melanin is very important dear  prevents cancer , 0.52 , 1.0 , pos
one word do what u like n love what you do , 0.5 , 0.6 , pos
flash light in front of us is clearly visible , 0.25 , 0.54 , pos
and also melanin is so important  it prevents cancer  and the chemicals which you used decreases melanin content and increase
s the probability of cancer , 0.4 , 1.0 , pos
don't believe them they all are giving paid reviews , 0.0 , 0.0 , neg
this is the kind audience base we need in india more  all the comments i see here are about being happy with our original ski
n color , 0.3 , 0.83 , pos
don't advertise these things , 0.0 , 0.0 , neg
harms others , 0.0 , 0.0 , neg
```

Fig. 6 Collected Data is stored in CSV file

D. Data Comprehension using Data Visualizations

From collecting polarity and subjectivity values from textblob.sentiment property, we observed the comments which were written against the use of glutathione product for skin whitening, or criticising the you tuber for endorsing the product is classified as Neg sentiment having polarity<0. Some of the comments rejecting use of product with language of acceptance were classified as Pos having neutral polarity , other comments which showed favour/ interest on trying the product were also classified as Pos with high levels of polarity. In other words sentiments which negative polarity <0.0 and neutral polarity <0.5 are definately against the use of glutathione , whereas comments with high polarity >0.5 are mostly in support of the you tuber endorsing the product and willing to use it.

```
In [408]:
import pandas as pd
df =pd.read_csv('Comments2.csv')
df.head()
```

Out[408]:

	text	polarity	subjectivity	pred class	actual class
0	clear skin is more important than lighter skin...	0.25	0.60	pos	against
1	i'm proud of my dark skin i don't want to be f...	0.54	0.82	pos	against
2	melanin is very important dear prevents cancer	0.52	1.00	pos	against
3	one word do what u like n love what you do	0.50	0.60	pos	against
4	flash light in front of us is clearly visible	0.25	0.54	pos	against

Out[304]:

	polarity	subjectivity
count	300.000000	300.000000
mean	0.161067	0.375800
std	0.326631	0.367314
min	-1.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.330000
75%	0.335000	0.670000
max	1.000000	1.000000

Fig. 7 Polarity and Subjectivity of individual comments

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

1) *Box plot:* The Boxplot is a graphical representation of the describe() method. From box plot we can understand the language used in comments are highly subjective as the topic of skin tone preference is personal to many Indians, also viewers preferred using neutral language to convey the message, outliers represent the Extreme conformity with product with most with harsh language used for criticism or Support for the you tuber such as fan base .

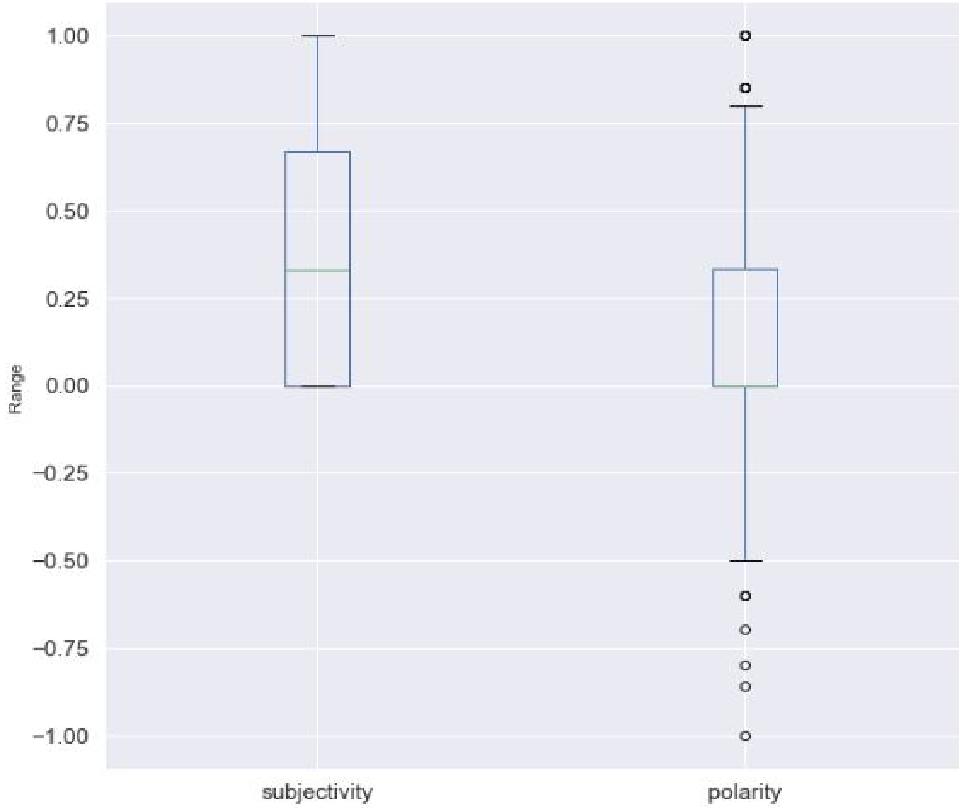


Fig. 8 Box plot of polarity and subjectivity

2) *Scatter plot:* From scatter plot we can understand the comments are highly subjective , meaning that viewers are heavly invested in the topic and are having strong sentiments to either support or reject the idea of skin lightening/use of glutathione . this is particularly true for high and low limits of polarity. Many comments have subjectivity as 1 , clearly depicting strong opinion on the topic. From the plot we understand most of the comments rejecting use of gluthathione of skin lightening are having strong opposition to the youtuber with high subjectivity and negative polarity.

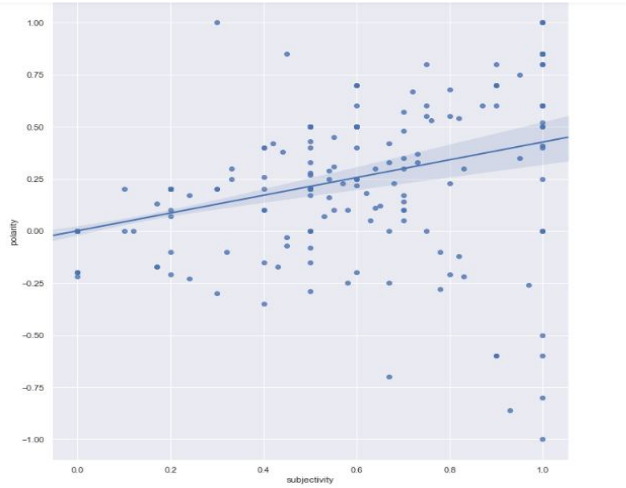
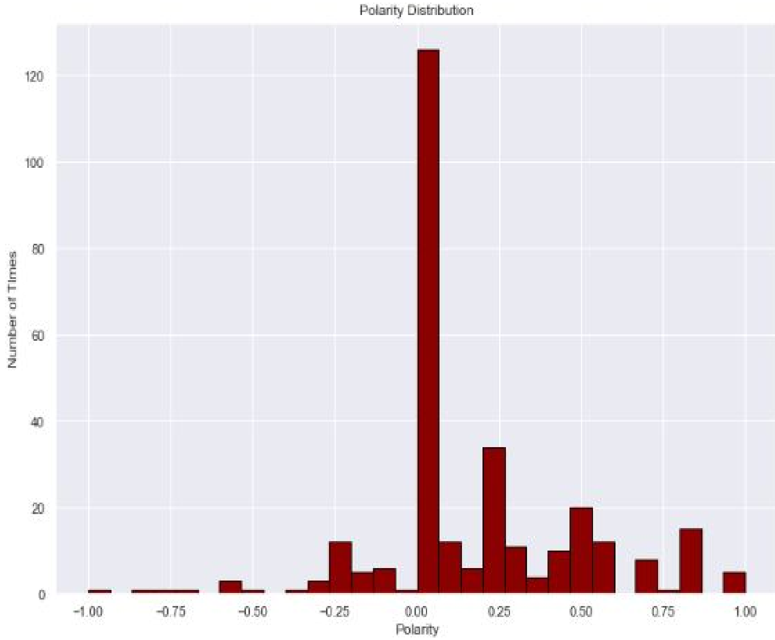


Fig. 9 Scatter plot of polarity vs subjectivity

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

3) *Polarity Distribution and Polarity Density Distribution*



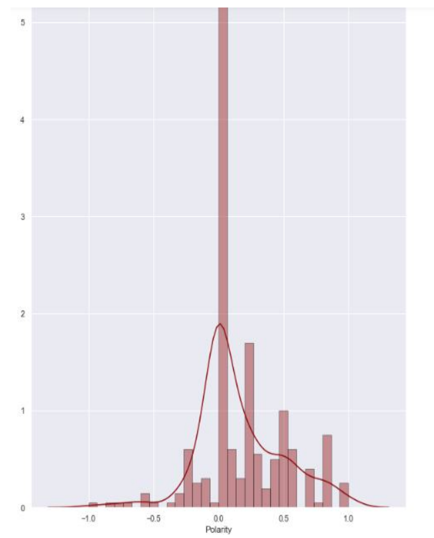


Fig. 10 Polarity Distribution on the left and Polarity Density Distribution plot on the right

From the Polarity Distribution plot we understand that the crowd is a mixed feeling crowd, all negative polarity distribution (harsh language) can be said for certain that the comments are against use of glutathione for skin lightening, most of the comments are using neural language are also not in support for glutathione whereas above 0.50 polarity are definitely interested in product inquiring about price, scars, duration, application and mostly are strong supporters of the YouTuber. We can understand there is almost equal distribution of crowd sentiments, slightly half of them are against the use of glutathione and the slightly lower half distribution are willing to try the product and interested in details such as price, brand, usage, effects on pimples and male skin etc.

4) *Frequent Words:* Frequent nouns used in the comments can be extracted on using Counter from collection which counts the occurrence of words from text which is processed to remove stopwords such as he,she,it,they etc and then counter on nouns using noun_phrases property in textblob given as below

```
from nltk.tokenize import word_tokenize
stopwords = nltk.corpus.stopwords.words('english')
word_tokens = word_tokenize(all_text2)
filtered_sentence = [w for w in word_tokens if not w in stopwords]
filtered_sentence = ''

for w in word_tokens:
    if w not in stopwords:
        filtered_sentence = filtered_sentence +' '+w

print(filtered_sentence)
from collections import Counter

nouns=TextBlob(filtered_sentence).noun_phrases
count_words = Counter(nouns)
total_words = len(words)
sorted_words = count_words.most_common(total_words)
```

Fig. 11 Collecting Frequent words data

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

Few of the most occurring words are then analysed to understand patterns

Out[320]:

	Word	Frequency
0	skin	50
1	dark	18
2	love	18
3	effects	17
4	beautiful	16
5	fair	14
6	happy	9

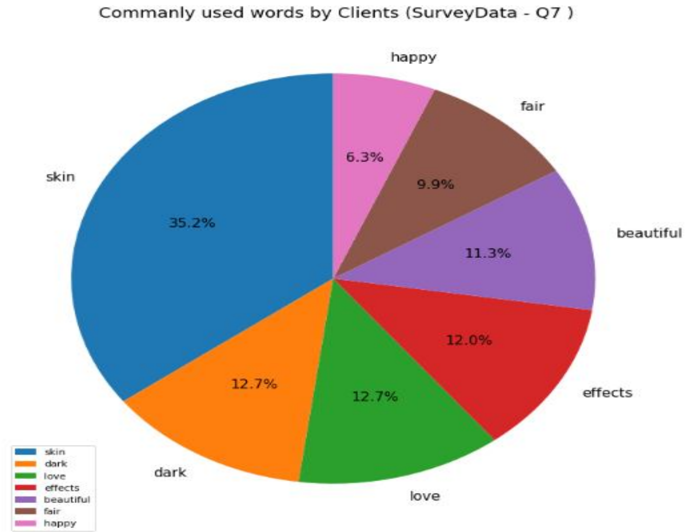


Fig. 12 Pie chart for Few Frequently occurring words

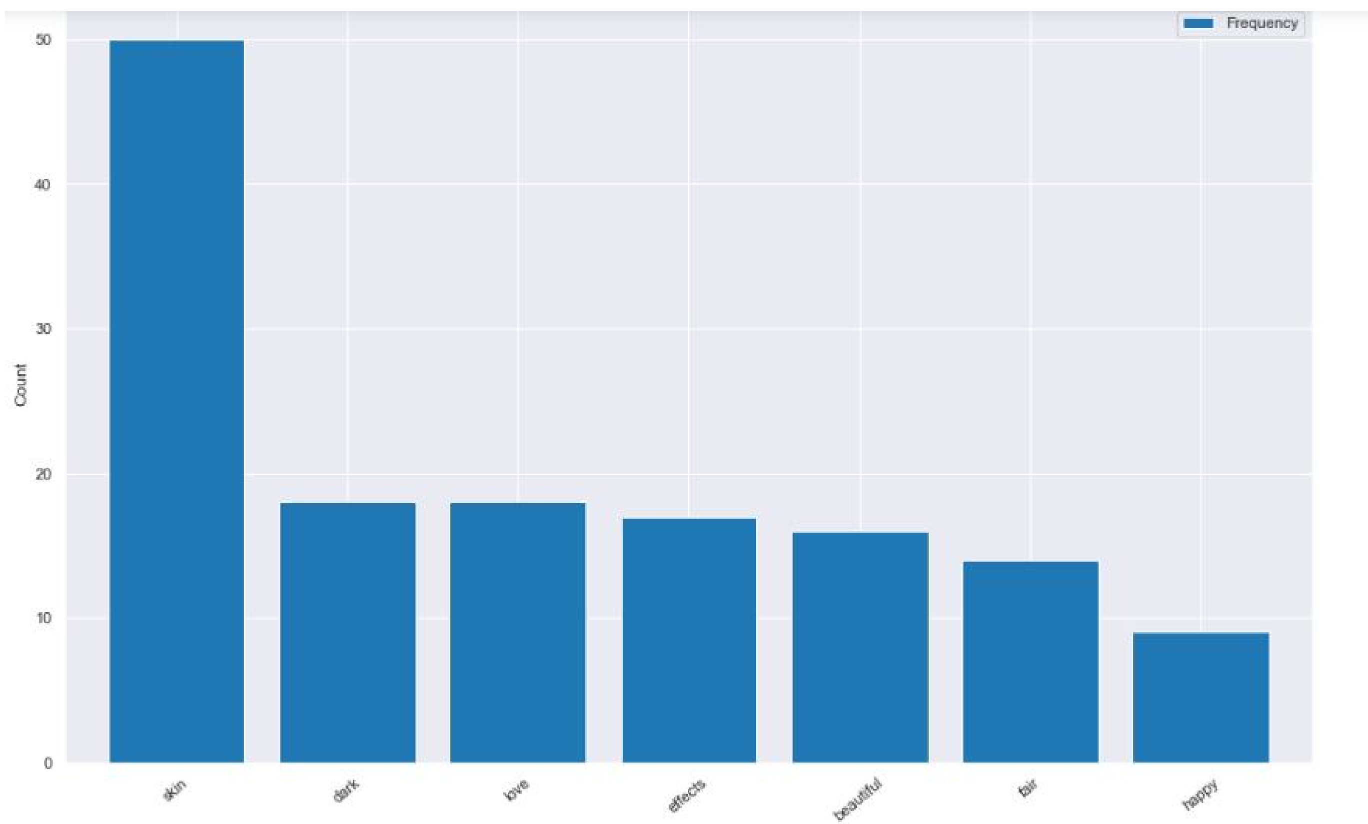


Fig. 13 Bar chart for Few Frequently occurring words

From the above charts we understand that viewers are speaking mostly about skin, with many keywords such as love, dark and happy used in comments against the use of product, also comments in favour of product also mention keyword such as side effects, pimples, price etc and were on the lines of appreciating the you tuber appearance. Interestingly promotion, money, paid were used to express dissent. The remaining frequently occurring keywords are as below

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

Fig. 14 word cloud of overall collected YouTube comments

III. CUSTOM CLASSIFIER FOR SENTIMENT ANALYSIS

The sentiment property of text blob gives good measure for polarity and subjectivity but accurate can be still improved, as it may not understand sarcasm, bias or preference precisely as a human being. This makes it essential to create custom classifiers for sentiment analysis that will classify newer comments much accurately then the sentiment property, which will help to understand change in opinions and trends on glutathione.

Also the class label generated as pos/neg based on polarity is objective based on English language usage, we have noticed there are quite an occurrence of misclassification of class based on incorrect generation of polarity. This makes it even more important to design custom curated sentiment analysis models which will more accurately classify the comments based on the topic of discussion and not purely on English language.

The custom classifier build is demonstrated in this section. We will demonstrate custom classifier using Naïve bayes algorithm and with the latest state of the art Deep neural networks provided by tensor flow.

A. Introduction to Naïve Bayes Algorithm

[6] Naive Bayes is among one of the most simple and powerful algorithms for classification based on Bayes' Theorem with an assumption of independence among predictors. Naive Bayes model is easy to build and particularly useful for very large data sets. The Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that a particular fruit is an apple or an orange or a banana and that is why it is known as "Naive".

1) *Bayes Theorem:* [6] Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It serves as a way to figure out conditional probability. Given a Hypothesis H and evidence E, Bayes' Theorem states that the relationship between the probability of Hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H|E) is :

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

This relates the probability of the hypothesis before getting the evidence P(H), to the probability of the hypothesis after getting the evidence, P(H|E). For this reason, is called the prior probability, while P(H|E) is called the posterior probability. The factor that relates the two, P(H|E) / P(E), is called the likelihood ratio. Using these terms, Bayes' theorem can be rephrased as: "The posterior probability equals the prior probability times the likelihood ratio."

Example of Bayes Theorem:

To find out the "Probability of the Card we picked at random to be a King given that it is a Face Card". So, according to Bayes Theorem, we can solve this problem. First, we need to find out the probability

P(King) which is 4/52 as there are 4 Kings in a Deck of Cards.

P(Face|King) is equal to 1 as all the Kings are face Cards.

P(Face) is equal to 12/52 as there are 3 Face Cards in a Suit of 13 cards and there are 4 Suits in total.

Then According to bayes theorem

P(King|Face) = (P(Face|King).P(king))/ P(King) = (1*1/13)/(3/13) =1/13

2) *The Game Prediction Problem:* Following Data, comprises of the Day, Outlook, Humidity, Wind Conditions and the final column being Play, which we have to predict.

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Fig. 15 Game Prediction dataset

First, we will create a frequency table using each attribute of the dataset.

Frequency Table		Play	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	3	2

Frequency Table		Play	
		Yes	No
Humidity	High	3	4
	Normal	6	1

Frequency Table		Play	
		Yes	No
Wind	Strong	6	2
	Weak	3	3

Fig. 16 Frequency Table of attribute

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

For each frequency table, we will generate a likelihood table.

$P(x|c) = P(Sunny|Yes) = 3/10 = 0.3$

Likelihood Table		Play: Yes	No	
Outlook	Sunny	3/10	2/4	5/14
	Overcast	4/10	0/4	4/14
	Rainy	3/10	2/4	5/14
		10/14	4/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 10/14 = 0.71$

Fig. 17 Likelihood table for outlook

Likelihood of 'Yes' given 'Sunny' is:
P(c|x) = P(Yes|Sunny) = P(Sunny|Yes)* P(Yes) / P(Sunny) = (0.3 x 0.71) /0.36 = 0.591
Similarly Likelihood of 'No' given 'Sunny' is:
P(c|x) = P(No|Sunny) = P(Sunny|No)* P(No) / P(Sunny) = (0.4 x 0.36) /0.36 = 0.40
Now, in the same way, we need to create the Likelihood Table for other attributes as well.

Likelihood table for Humidity

Likelihood Table		Play: Yes	No	
Humidity	High	3/9	4/5	7/14
	Normal	6/9	1/5	7/14
		9/14	5/14	

Likelihood table for Wind

Likelihood Table		Play: Yes	No	
Wind	Weak	6/9	2/5	8/14
	Strong	3/9	3/5	6/14
		9/14	5/14	

$P(Yes|High) = 0.33 \times 0.6 / 0.5 = 0.42$

$P(No|High) = 0.8 \times 0.36 / 0.5 = 0.58$

$P(Yes|Weak) = 0.67 \times 0.64 / 0.57 = 0.75$

$P(No|Weak) = 0.4 \times 0.36 / 0.57 = 0.25$

Fig. 18 Likelihood table for humidity and Wind

Suppose we have a Day with the following values :
Outlook = Rain
Humidity = High
Wind = Weak
Play =? (Output variable)
So, with the data, we have to predict whether "we can play on that day or not".
Likelihood of 'Yes' on that Day = P(Outlook = Rain|Yes)*P(Humidity= High|Yes)* P(Wind= Weak|Yes)*P(Yes)
= 2/9 * 3/9 * 6/9 * 9/14 = 0.0199
Likelihood of 'No' on that Day = P(Outlook = Rain|No)*P(Humidity= High|No)* P(Wind= Weak|No)*P(No)
= 2/5 * 4/5 * 2/5 * 5/14 = 0.0166
Now we normalize the values, then
P(Yes) = 0.0199 / (0.0199+ 0.0166) = 0.55
P(No) = 0.0166 / (0.0199+ 0.0166) = 0.45
Naïve bayes model would predicts that there is a 55% chance there will be a Game tomorrow.

B. Implementation of Naïve bayes custom model for Sentiment Analysis

Documentation: https://textblob.readthedocs.io/en/dev/classifiers.html#loading-data-and-creating-a-classifier

The textblob. classifiers module makes it simple to create custom classifiers. It uses the concept of Naïve bayes internally to generate model based on train data set . Note Accuracy of the model can be improved significantly by using larger and quality dataset generated from youtube comments. Code for training Model and using it for classifying is as below

```
df.head()
train = list(zip(df['text'],df['actual class']))

from textblob.classifiers import NaiveBayesClassifier
cl = NaiveBayesClassifier(train)

cl.classify("don't advertise")
cl.show_informative_features(5)

prob_dist = cl.prob_classify("melanin is very important dear  prevents cancer")
print(prob_dist.max())
print(round(prob_dist.prob("for"), 2))
print(round(prob_dist.prob("against"), 2))

cl.accuracy(test)
```

Fig. 19 Naives bayes classifier provided by Textblob

C. Custom sentiment Models with Newer Technologies such as TF-Hub

Documentation: https://www.tensorflow.org/hub/tutorials/text_classification_with_tf_hub

1) Installation

Necessary libraries can be installed as below

>>pip install -q tensorflow==2.1.0

>>pip install -q tensorflow-hub

>>pip install -q seaborn

2) Preparation of DataFrame for Training Model: The dataframe should contain columns pertaining to the individual text and corresponding polarity, in TF-Hub we should keep positive polarity as 1, and negative polarity as 0. Example of dataframe format is as below

```
In [137]:  df =pd.read_csv('SampleData.csv')

In [139]:  df.head(10)
Out[139]:
```

	text	polarity	subjectivity
0	clear skin is more important than lighter skin...	1	0.60
1	i'm proud of my dark skin i don't want to be f...	1	0.82
2	melanin is very important dear prevents cancer	1	1.00
3	one word do what u like n love what you do	1	0.60
4	flash light in front of us is clearly visible	0	0.54
5	and also melanin is so important it prevents ...	0	1.00
6	don't believe them they all are giving paid re...	0	0.00
7	this is the kind audience base we need in indi...	0	0.83
8	don't advertise these things	0	0.00
9	harms others	0	0.00

Fig. 20 DataFrame format for TF-Hub

Then we use Estimator framework provides input functions that wrap Pandas dataframes. This creates train_input_fn and predict_train_input_fn which will be used to training and evaluating dataset.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

TF-Hub provides a feature column that applies a module on the given text feature and passes further the outputs of the module. The code uses nnlm-en-dim128 module. In code embedded_text_feature_column represents the feature column. The module specified in embedded_text_feature_column is responsible for preprocessing of sentences (e.g. removal of punctuation and splitting on spaces) given as module_spec property.

For classification model we use Deep neural network classifier , instance name is estimator as given below

```
In [140]:
# Training input on the whole training set with no limit on training epochs.
train_input_fn = tf.compat.v1.estimator.inputs.pandas_input_fn(
    df, df["polarity"], num_epochs=None, shuffle=True)

# Prediction on the whole training set.
predict_train_input_fn = tf.compat.v1.estimator.inputs.pandas_input_fn(
    df, df["polarity"], shuffle=False)

In [141]:
embedded_text_feature_column = hub.text_embedding_column(
    key="text",
    module_spec="https://tfhub.dev/google/nnlm-en-dim128/1")

In [142]:
estimator = tf.estimator.DNNClassifier(
    hidden_units=[500, 100],
    feature_columns=[embedded_text_feature_column],
    n_classes=2,
    optimizer=tf.keras.optimizers.Adagrad(lr=0.003))

INFO:tensorflow:Using default config.
INFO:tensorflow:Using default config.
WARNING:tensorflow:Using temporary folder as model directory: C:\Users\mosa0816\AppData\Local\Temp\1\tmp9ocu058h
WARNING:tensorflow:Using temporary folder as model directory: C:\Users\mosa0816\AppData\Local\Temp\1\tmp9ocu058h
INFO:tensorflow:Using config: {'_model_dir': 'C:\\Users\\mosa0816\\AppData\\Local\\Temp\\1\\tmp9ocu058h', '_tf_random_seed': No
ne, '_save_summary_steps': 100, '_save_checkpoints_steps': None, '_save_checkpoints_secs': 600, '_session_config': allow_soft_p
lacement: true
graph_options {
  rewrite_options {
    meta_optimizer_iterations: ONE
  }
```

Fig. 21 DNN Calssifier

3) *Train and Evaluate the Model:* The below code is used to train, evaluate the accuracy and display the confusion matrix, in our case we got mediocre accuracy for demonstration purpose, which can be improved using more quality dataset and hyper tunning.

```
estimator.train(input_fn=train_input_fn, steps=5000);
train_eval_result = estimator.evaluate(input_fn=predict_train_input_fn)
print("Training set accuracy: {accuracy}".format(**train_eval_result))
```

Fig. 22 Training model code

IV. RESULTS AND CONCLUSION

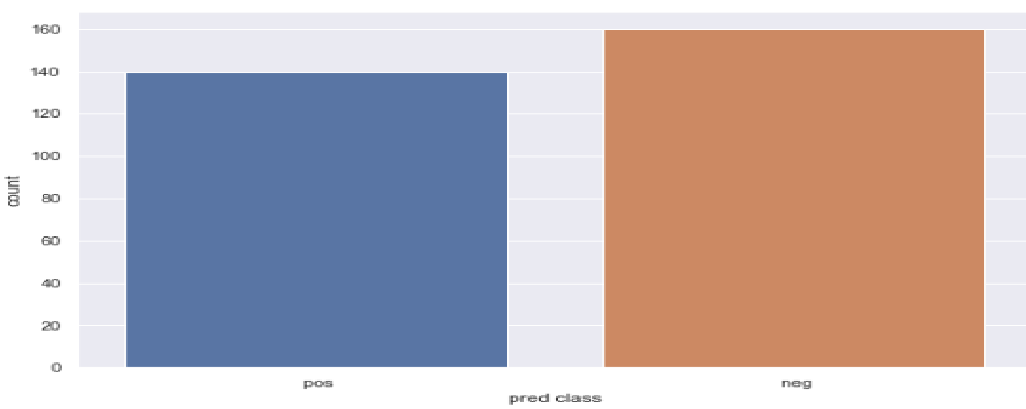


Fig. 23 Countplot of Pos vs Neg comments, Neg comment represent against used of product, few positive comments are against the use but 90% are in favour of use of glutathione of skin whitening

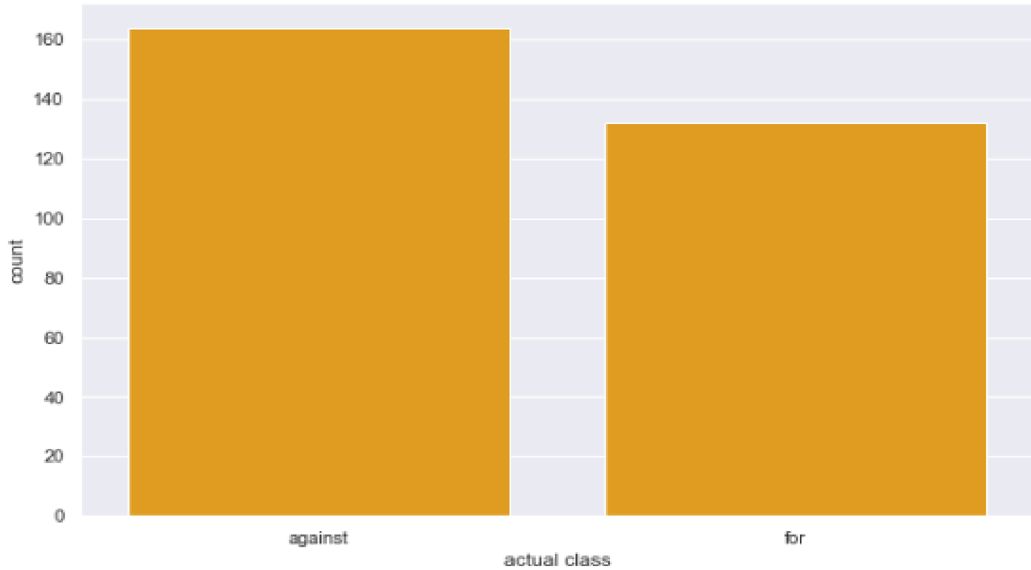


Fig. 24 Countplot of comments against the use and for the use of glutathione for skin whitening

Using various techniques primarily from Selenium framework for YouTube Comment extraction and Text Blob API for sentiment analysis , we understand the sentiment of Indian viewers on video endorsing glutathione product for skin whitening, we see equal distribution of sentiments , more than half of the entire comments rejected the idea of skin whitening products with harsh criticism or with sarcasm, also critics against the product also personally attacked the you tuber as a way for rejection . Overall Subjectivity of comments was very high, probably due to the reason being topic of skin lightening being a personal topic to Indians so there was lot of display of strong emotions of rejection. Unfortunately the other half of the comments depicted that they were interested in trying the product with many inquires about Pricing, duration, application on boys and potential reduction in pimples. Also comments in favor showed the few of the viewers fully supported the you tuber and were easily perusable by the you tuber. These insights gives hopes that skin tone preference and endorsing skin lightening products would be strongly opposed in near future, also we can clearly see a shift to negative outlook to using skin lightening product such as glutathione and strong acceptance to colorism is a good shift for the society.

Few plausible Reasons for Shift in sentiments :

1) Beauty standards have changed globally especially in modelling business, there are many successful models that have changed the perspective of beauty and enriched the concept of self acceptance, few of them are Diandra Forrest having albino, Winnie Harlow having Vitiligo, Madeline Stuart having down syndrome ,Tess Holliday plus size model and many more have changed the standards of beauty. And Indians are very much aware of these through online content.
2) Many dusky and darker tone women have been crowned the miss world title like 6 times from India, 6 times from Venezuela , 4 times from Jamaica etc , which is changing the global mindset of beauty.
3) Representation of people of different races and skin tones currently being casted as roles in Hollywood movies which has also started to impact the minds of global audiences to accept diversity. Bollywood is also adapting to the global change by making more impactful movies such as chapak, baala etc which are starting to break stereotypes but has a long way to go.
4) Combating preference of lighter skin tone with the awareness of Science of Melanin such as it protects the skin from uv rays preventing skin cancer, acts as a natural makeup to the skin hiding scars, research study on how melanin skin ages slower than pale counterparts and hides fine lines.

REFERENCES

[1] Savita Malik, “Domination of fair skin: Skin whitening, Indian women and public health” November 2010 Conference: 138st APHA Annual Meeting and Exposition 2010
[2] Dilan K Jayasekara , https://towardsdatascience.com/statistical-sentiment-analysis-for-survey-data-using-python-9c824ef0c9b0
[3] https://www.tensorflow.org/hub/tutorials/text_classification_with_tf_hub
[4] https://textblob.readthedocs.io/en/dev/classifiers.html#loading-data-and-creating-a-classifier
[5] Kislay Keshari, https://www.edureka.co/blog/naive-bayes-tutorial/