**YouTube Comments analysis using NLP – A Case Study**

In today’s fast moving and data driven world, it is essential for corporates to gather and to know as much information about their potential and current clients as possible. Each trivial step could boost some company’s turnover in a roll over effect that is now becoming difficult to predict. The details of our work schedule, our hobbies, are sentiment, our preferences, everything can be tracked, analysed, modelled and then monetized. Large corporations all over the world are harnessing the power and acumen of ever improving machine learning algorithms to create a product to maximize their revenue by channelling our data and preferences over their models. One such technique is **sentiment analysis**.

As explained by Wikipedia, Sentiment analysis or opinion mining is a field in data science refers to the use of Natural Language Processing (NLP), text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. We use sentiment analysis to voice the important quantifying elements of a customers’ interaction with a system such as, reviews, responses, surveys, comments. In other words, here we try to quantify and analyse users' ideas, beliefs, opinions, and much more to construct a clear picture that can help our pre-defined problem statement in the fields of marketing, research, politics, surveys etc.

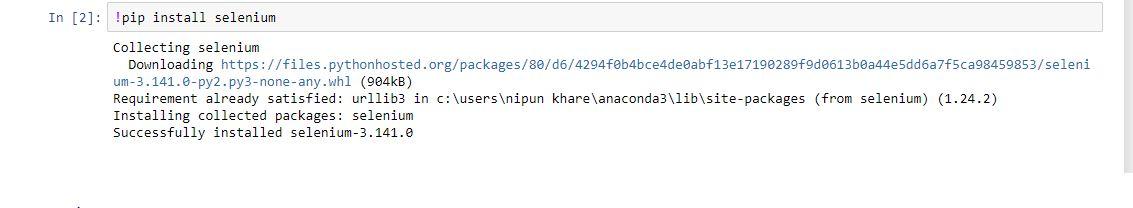
But before we define our problem statement, we need to know a little more about **NLP** itself. Thus, as per Wikipedia, Natural language processing or (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human languages. In particular how to program computers to process and analyse large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

So now we can move forward with our problem statement, or what we’re trying to achieve through this project. Here we want to study and analyse the latest event of farmers’ unrest and agitation in Delhi and in many parts of India due to the three new **Farms Acts**. According to the Union Government, these three new laws would revolutionize the farming sector in India and would massively benefit the farmers of our nation. These could be a crucial step towards doubling the income of farmers by 2022 – a goal set by the government a few years ago. But the other side of the story seems to be a little different, many farmers and farmers organizations claim that these new acts would jeopardize their livelihood and the basic support system from the government which will be fatal for the small and marginal farmers of India who make 86% of their population.

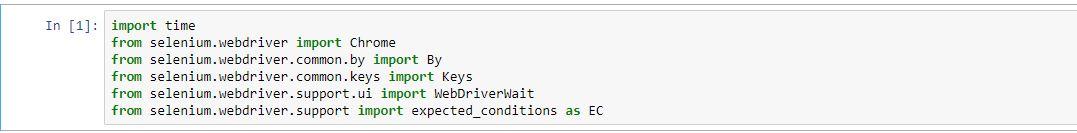
Here, we won’t try to analyse these acts itself but the public’s perception and opinion on them. These kinds of studies could be useful for government organisations and their think tanks to try to evaluate what people think about their actions that affects society at large.

For analysing the public’s perception towards these 3 new acts, we’ll be using **YouTube comments**. Throughout the farmers' protest against these laws, there were numerous trending videos that tried their best to encapsulate the issue as thoroughly as possible. We’ll be taking some of those trending videos for our analysis with a set criteria like maximum views within a particulate time period and also with maximum interaction in the comments section.

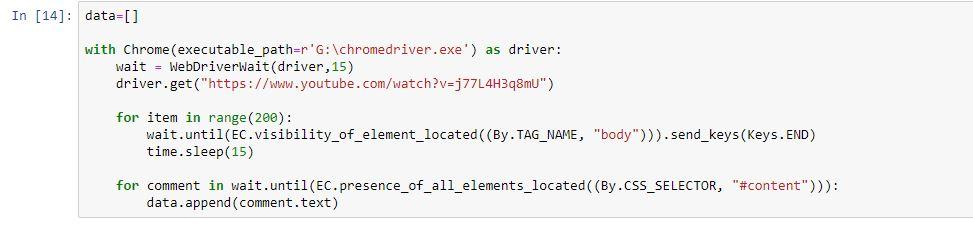
Our work starts with **scrapping** the comments of trending videos of farmers protest. First, we need to import required packages for the given task, starting with **selenium**.



The selenium package is used to automate web browser interaction from Python, this package would help us to automate the **web scrapping** of YouTube comments from the trending videos. Next step is to import the necessary libraries –



Now we can start scrapping for top comments from the top videos across the platform. We can use the following code to get the job done.



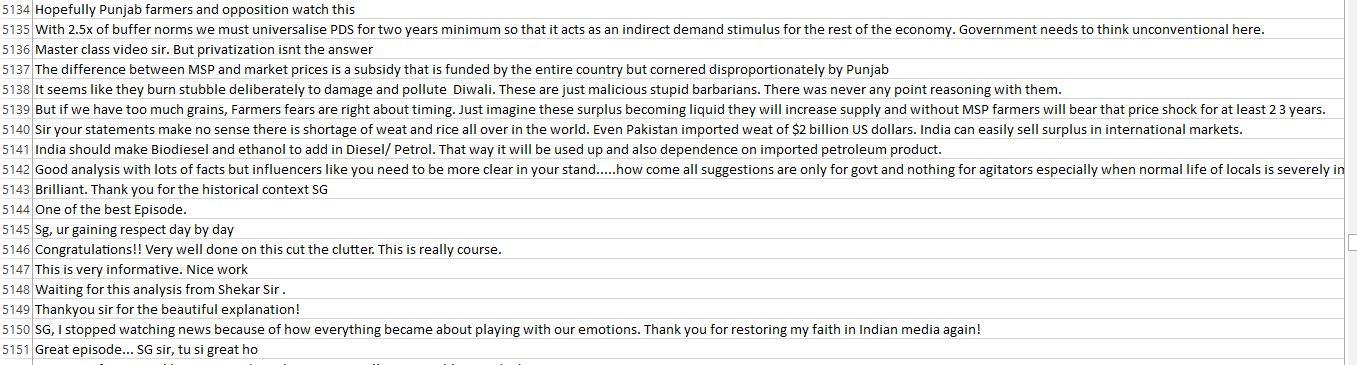
Using the Selenium **web-driver**, we scrapped the comments. Next, we need to make a data frame out of the scrapped comments, and for this we use the **Pandas** library of python. Result of the scrapped comment is shown below.



Repeating this process to scrap the comment over all the selected videos would generate us a corpus of variety of comment, both in favour and against the farmers' protest. We used the code over 12 trending videos and generated a basket of over **7000 + comments**. But for working with our gathered dataset, we first need to download all the scrapped comments using the following code –



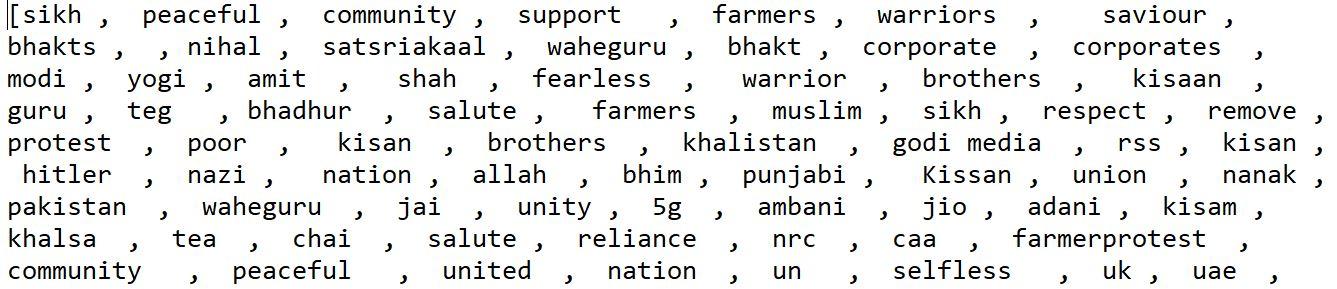
Making 12 such excel files and then combining them all together would get us our final combined dataset. A glimpse of the corpus comments –



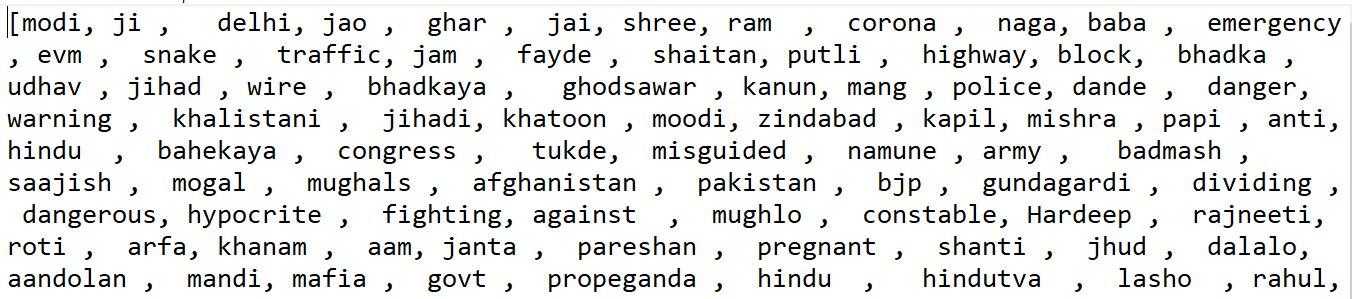
In our downloaded dataset, there were lots of emojis, text in Hindi, Punjabi and Bengali language, some symbols, empty spaces and lots of interfering matter that could hamper the end result. Thus the next order of our business was to **clean the data** and remove all the unwarranted elements from it and then apply NLP models to it.



The job begins with importing our combined dataset in the python workbook. Then in order to classify a given text either in favour or against the farmer’s agitation and protest, we need to custom build a **dictionary** of relevant for and against words from our combined datasets. We carefully went through and examined each and every comment to extract a workable custom dictionary that suites our problem statement. In process, we created two **1000 +words dictionaries** to correctly flag a comment either in favour or against the farmer’s unrest. One named ‘anti’ and one ‘support’. A small part of the custom dictionary for comments in favour of the unrest –

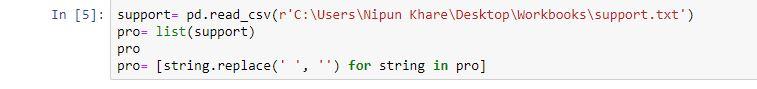


And for those comments who oppose the said agitation –



Next step was to embed the custom dictionary in the workbook and the dataset and convert its type from Dataframe to list, using the following code. And in the process to clean some irregularities in them.





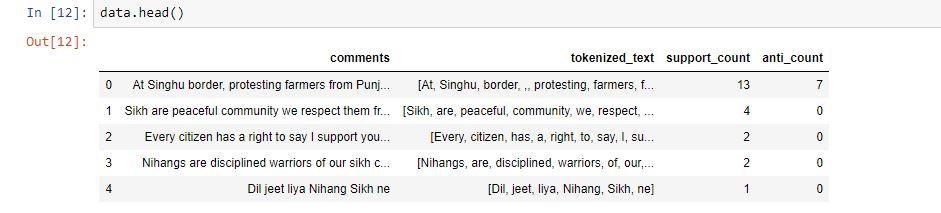
Next step in the process is to **tokenize** the scrapped comments using the Tokenize library from **NLTK**. It is the process of breaking down the given text or paragraphs into smaller chunks such as words or sentences. These tokens are the core building block of the sentences and phrases we use in our daily life. When we apply the code, we get the following output –



Tokenized text is easy to work with and makes the task of appending the dictionary to classify the sentiment a simple process. The next step in our process is to create a function that goes through all of our combined comments dataset and classify the text according to the condition specified in the code. The loop goes through each and every comment and tags the tokenized words in it according to the custom dictionaries and then counts the total words in support or against the motion in a given comment. The mode of the count decides the sentiment of a given comment.

The code and the output is as follow –



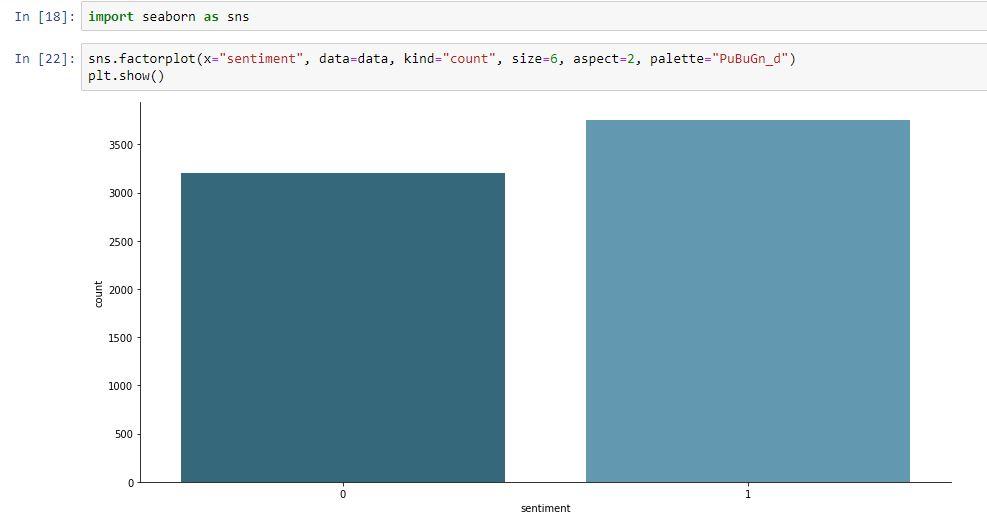


The **classification** of a given comment as in support or against the farmer’s agitation is determined by the majority word count of a distinct word type as specified in the dictionaries. In case of a tie in the number of words in support or against the protest, as the condition laid out, the algorithm classifies those particular comments as against the protest. (This condition can be changed easily through the function code)

The code and the output is as follow-



Now, we have our required result, sentiment of all the comments of the most trending YouTube videos over a particular matter. But it’s still not easy to grasp the full picture of our result just by a mere count. To paint a more clear picture of the situation here, we can use the **visualization** tools to see the actual number of comments in favour or in against the protest using a simple graph using the **seaborn** library, as following –



Now, it’s much clearer that even with our few biased conditions, the majority of the comments support the farmer’s protest and that’s something the concerned authorities should take a note of.

**Summary -**

Machine learning tools like NLP are boon for society. When used in a constructive manner and not only for monetary gains, it can reveal and de-cluster a lot of sentiments of society as a whole. And we tried to do the same through this project. Here, we analysed the latest farmer’s unrest in the country by the means of YouTube comments of trending videos on the subject. Though the study is not compressive enough, it gives us a near accurate outlook at people’s perception towards this issue. We gathered enough information to state some findings like the political matters like these are becoming much more polarizing than we might like to accept. People from one group are more or less adamant to prove the ‘others’ wrong. There is hardly any content or media house over the internet that is not trying to paint a particular colour to these historic protests. The interpretation of these acts swings left and right based upon who is asked. This is proven by the fact that the videos supporting the cause have most of the views from the people who share the same sentiment and same goes for those who are against these protests. The ‘other’ side’s interpretation is always wrong and there is hardly any sympathy between those who share differential views. We tried to be as impartial in the process as possible but still there might be some distortion in the results due to biases from unforeseen circumstances as stated in the conclusion as well. Thus, these results of our findings should be taken with a bit of a grain of salt. But as far as the final output goes, one interpretation of the results might be that, majority of people in the country supports the farmers’ cause and are sympathetic towards it. They want the Union government to take appropriate action to end the farmer’s misery and blame their inefficiency for the state of the matter as of now.

Needless to say, we can delve deeper into this case study and find out more with the tools available with us using NLP. The scope of study like this is broad with great ramifications and this should be considered incomplete as far as the spirit of true analysis goes, but sufficient enough for the task in hand.

**Conclusion-**

Through the analysis of over 7000 comments on trending YouTube videos over the latest farmers’ protest, we gathered that the majority of people in the public forum are **supporting the farmers’ cause** and are sympathetic towards it. There is a clear verdict that people want the Union government to take some immediate action to put farmers out of their misery. There might be some biases involved in the final output by the mere facts like – the videos were handpicked, the dictionaries made were as perfect as our due diligence, and the code to classify the comment in case of a tie between keywords are all based on our input. The study here demonstrates the working principles and applications of Natural Language Processing and results like these could be used by the concerned authorities to pan out a more detailed outlook of public perception towards a particular issue.

The study here is not comprehensive enough but the future scope of such a study is needless to say impressive. We can gain more insight into the matter by analysing the underline 3 legislation itself, its key provisions and key points to get ourselves in a better and informed position.