# Contribution of Competitor’s Article

I chose air pollution as a data science problem, the main objective is to measure the AQI using MLP and Uses Text Processing to Process the text form the Social Media which produces near real time forecast. Each Pollution have various effect on the living organisms and the environment such as environmental degradation, global warming, ozone depletion, climate change and these are lot of others but these are some of the major effects. Air contamination is a critical issue in China and somewhere else around the globe. For instance, in 2013 Beijing had 58 days at the point when the Air Quality Index (AQI) was higher than 200 or substantial pollution. In December 2013 the east and focal locales of China, which have more than 600 million individuals, experienced substantial contamination for over two weeks. Air contamination is unsafe to individuals' wellbeing, bringing on eye bothering, lung and throat disturbance, lung growth and issues with infants at birth To better manage the issues of air contamination, the first step is to screen air quality. From January 1 to November 1, 2013, the scope of physical observing stations has expanded from 74 urban communities to 108 urban communities in China. Likewise, the Chinese government has begun to incorporate PM2.5 (a noteworthy what's more, unsafe air contamination) into AQI monitoring. The cost of setting up and keeping up physical observing stations restricts their arrangement at present to substantial and medium urban areas as it were. Subsequently, AQI checking in numerous areas, for example, little urban areas and country towns is as yet deficient. To help individuals in these districts get air quality data, we consider the accompanying inquiry: would we be able to evaluate AQI without physical checking by utilizing other, officially accessible, data sources? In this paper we assess AQI utilizing online networking information as the data source. Online networking is a rich and convenient data source about air contamination in China. The calculated weight of the word haze on the social media post is positively correlated with AQI Markov Random Field (MRF) model to model the correlation between AQI and social media post. KNN is used to estimate the AQI for the nearby cities from where the social media post were collected.

# Description of Your Contribution

For this Project we will be using two methods Natural Language Text Processing for Text Processing and Multi Layered Perceptron a recurrent neural network. MLP has three units input gate, output gate and an activation function. This algorithm is based on the boosting algorithm where difficult points of the time series are concentrated on during the learning process however, unlike the original algorithm; we introduce a new parameter for tuning the boosting influence on available examples. We test our boosting algorithm for RNNs on single-step-ahead and multi-step-ahead prediction problems which anticipate the air quality one day ahead of time for the locations that are provided in this dataset and also by splitting the whole dataset into training and testing data to evaluate the accuracy of the prediction. Location data will be applied to deal with coverage limitation of air pollution monitoring sensors. Reduce number of sensors by predicting the air quality index using means of reading social media. Because social media is prevalent in the population it produces a lot of data which can be collected and analyzed.

# Data Source and Content

City Pulse Air pollution dataset which has the values for carbon monoxide, nitrogen dioxide, sulfur dioxide, particulate matter and ozone index levels. The measurements of these Pollutants are collected from August 2014 to October 2014 from a 449 sensors City of Brasov in Romania. AQI Prediction using the selected Bag of the words Vector

The air quality index is a way to calculate the air pollution level. In order to calculate the Air Quality Index we need to measure the level of five major pollutants in the atmosphere carbon monoxide, nitrogen dioxide, sulfur dioxide, particulate and ozone index level. The value of the air quality index ranges from to greater than 100 and it is split in to 5 categories 0 to 25 is considered as very low pollution, 25 to 50 is considered as low pollution, 50 to 75 is considered as medium pollution, 75 to 10 is high and values above 100 is very high said to high air pollution and it is extremely hazardous to health. These categories are color coded to communicate in an easy way to the people very low as green, low as light green, Medium as yellow, high as orange and very high as red.

Using Social Media Post for a one-month period from different cities to predict the Air Quality Index. In Text processing the documents are represented into a bag-of-words vector. Collected tweets using Twitter API that are posted from New York City about air pollution for a 10 day time period.

# Your Method

The MLP consist of one hidden layer and 14 neurons on the hidden layer. Neural systems are nonlinear models used to surmised answers for complex issues and can be utilized to display any nonlinear capacity. They procure information of the framework or condition they are inserted in through perceptions and utilize them to prepare the system. Repetitive neural systems (RNNs) are dynamical frameworks that are particularly intended for fleeting issues, as they have both criticisms as well as encourage forward associations. The general structure of a RNN comprises of synaptic associations between the information, covered up and yield layers of neurons. Learning is spoken to in a system by the estimations of these synaptic associations. The conditions of the neurons are subject to these free parameters, the contributions to the neurons and the states of the neurons at past time steps. A RNN can have duplicates of any neuron in the 80 arrange from the past time-step and they can be utilized to impact the expectation of information at future emphases. The goal of learning is to prepare the system by conforming the association weight values, more than a few preparing ages, to decrease the yield mistake of the system. Preparing moves the blunder towards a base point on the mistake surface, which has the free parameters of the system as its directions

Created a document vector for each document and the dimension of the vectors will be the number of distinct terms in the Bag of Words. Calculated the absolute frequency of a term for each bag of words that created for each tweets

The feature extraction method, extracts the aspect (adjective) from the dataset. Later this adjective is used to show the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using unigram model. Unigram model extracts the adjective and segregates it. It discards the preceding and successive word occurring with the adjective in the sentences. For above example, i.e. “painting Beautiful” through unigram model, only Beautiful is extracted from the sentence. Support vector machine examines the information, characterize the choice limits and uses the bits for calculation which are performed in info space. The info information are two sets of vectors of size m each. At that point each information spoke to as a vector is grouped in a specific class. Presently the errand is to discover an edge between two classes that is a long way from any archive. The separation characterizes the edge of the classifier, amplifying the edge diminishes uncertain choices. SVM likewise bolsters arrangement and relapse which are valuable for measurable learning hypothesis and it helps perceiving the figures accurately, that should be considered, to comprehend it effectively.

After the preparation and characterization we utilized semantic investigation. Semantic investigation is gotten from the WordNet database where each term is related with each other. This database is of English words which are connected together. On the off chance that two words are near each other, they are semantically comparative. All the more particularly, we can decide equivalent word like closeness. We delineate and analyze their relationship in the metaphysics. The key errand is to utilize the put away records that contain terms and after that check the comparability with the words that the client utilizes as a part of their sentences. Along these lines it is useful to appear the extremity of the opinion for the clients.

# Quantitative Results 1

Once the model is fit, we can appraise the execution of the model on the train and test datasets. This will give us a state of correlation for new models. Take note of that we alter the expectations before ascertaining blunder scores to guarantee that execution is accounted for in an indistinguishable units from the first information At last, we can forecasts utilizing the model for both the prepare and test dataset, got a root mean square error estimation R square is 0.86, MSE 0.05, Mean Absolute error 0.06 which is not actually not a bad model considering the input size.

# Quantitative Results 2

My other part of the project includes mining social media data to predict the air pollution. Mesuring the AQI using Text Processing to Process the text form the Social Media can produce near real time. Data collection is done by searching twitter to collect data with various hash-tags like #Pollution, #airquality #airpollution etc for same dates to that of sensor dates and also for the same location. Organized Tweets are by and large in sentence design, with URLs indicated for pictures or blog articles. To get information that is in usable configuration we evacuate the stop words that contains general terms like a, the, and so on and emoticons. We performed taking after operations on tweets in purifying and normalizing stage. We utilize the unigram, bigram and a Unigram and Bigram (half and half) highlight extraction strategy for study reason. Crossover elements are taken for outright positive words like "brilliant", "amazing", "dependably" and so forth and negative words, for example, "never", "not", "scarcely" and so on. To separate the amplified highlights, we utilize the changed demonstrate. Any term that happens least of 500 circumstances has been taken as an amplified target. We utilize K =20 in extricating top K things from the terms that showed up for more than the limit esteem. This technique helps getting related terms that can be mapped with unique target questions. The improved dataset after pre- processing has a lot of distinctive properties.

# Discussion: Comparison with Your Competitor

The methods that I used are AQI Prediction using the selected Bag of the words Vector, Normalized the AQI to train the neural network, AQI Prediction using Multi Layered Perceptron and to that of my competitors were AQI Prediction with the full Bag of Words vector , AQI Prediction with K-Nearest Neighbor, AQI Prediction with Markov Random Field

# Performance on Big Data: Time Measurements

Text Preprocessing took around 18 seconds, Multilayered Perceptron 2 minutes, Bag of Words Creator 5 minutes.

# Conclusion

Analyzing the social media data may be a solution to reducing the cost involved in deploying the sensors to monitor air pollution in cities plagued by an increasing rate of pollution. They can be more useful in the places where there are no monitoring stations to estimate the air quality. Social media may be a solution to reducing the cost involved in deploying the sensors to monitor air pollution in cities plagued by an increasing rate of pollution. This will give the government a long term solution utilizing this method. Furthermore to come up with an estimator for the social media data that are used for analyzing this Project may be able to provide a novel solution to an increasing concern for society and the health of the population.