Semi-Supervised Classification of Climate Change & Population Related Tweets based on Hash Tags and Accounts annotation

Chongtham Rajen Singh[1], Dr. R. Gobinath[2], and Dr. Nongmaithem Ajith Singh[3]
[1]*Ph.D. Research Scholar, Department of Computer Science, VISTAS, Chennai, India*
[2]*Associate Professor, Department of Computer Science, VISTAS, Chennai, India*
[3]*Assistant Professor, Department of Computer Science, South East Manipur College, Manipur, India*

***Abstract* — Many researchers work on climate change related tweets to predict and determine whether climate change is real or hoax based on sentiment analysis using labeled dataset. Others work on predefined climate change denier hash tags or denier twitter accounts to annotate these tweets. This paper illustrates the combine usages of denier hash tags and predefined twitter accounts that are primarily denier of climate change propaganda in Twitter. Denier accounts and hash tags collected from across various papers, articles are used along with additional wild character search techniques of seed phrases related to climate change and population growth related words or phrases together and annotated the unlabeled dataset extracted from Twitter. It is an automatic annotation technique. This annotated subset of data is used to train baseline Supervised Classification Models in combination with two types of frequency-based word vectorizers to analyze the performance measure of each model with different feature variation of n-gram combination. As per the analysis Linear Support Vector Machine algorithm along with word Count Vectorizer of Unigram and Bigram combination score the best performance on the annotated dataset that is being used.**

***Index Terms* — Climate, Population, Denier, Believer, Annotation, Hash Tag, Account, Vectorizer, Supervised Classification**

I. INTRODUCTION

Unsupervised Machine learning (such as K-Mean clustering, Word Dependencies and Rule Based) has shown to perform poorly for the twitter dataset that is being used in this research paper. This dataset is a collection of *climate change and population growth* related tweets for the last 10 years (2009-2019). Refer [4]. As per a research article, *rich countries believe that climate change is caused due to increase in human population* and *poor countries believe that the assumed hypothesis is incorrect*. (TOPRO: The Overpopulation Project). It is motivated by this assertion, and this research paper is a preliminary step towards proving the hypothesis that has been stated. This unlabeled dataset cannot simply go by sentiment classification using unsupervised algorithms. Sentiment analysis aims at identifying the emotions, feelings and subjectivity of the opinion holder and will not hold correct when trying to classify tweets as *denier* or *believer* based on a topic or target entity. Each denier or believer tweets will have either Positive or Negative or Neutral Sentiment. The opinion holder of such two categories i.e., *'Denier'* and '*Believer'* holds separate sentiment for each category. Therefore, some sort of stance detection or automatic classification of the *climate change and population* related tweets into Believer & Denier is essential to prove the hypothesis. Manual annotation of the dataset consumes time and money. This is one of the disadvantages of Unsupervised Machine Learning. It performs poorly in the areas of problem solving, planning and decision-making task.

This research paper presents a combine technique of annotating the unlabeled dataset automatically using *denier and believer hashtags, denier twitter accounts, and seed word phrases related to climate change and population keywords.* Not all the tweets in the unlabeled dataset can be annotated using this technique; however, some subset of the tweets can be categorized as Believer & Denier. These labeled data have been used to train baseline Supervised Learning Algorithms (such as Logistic Regression, Linear SVM, Naïve Bayes, Decision Tree) along with frequency-based count vectorizers (Count Vectorizer, TF-IDF Vectorizer) of unigram, bigram and trigram to pick the best performing model for further prediction and analysis of the remaining unlabeled dataset. It is a Semi-Supervised Machine Learning Technique.

II. LITERATURE REVIEW

An opinion holder may deny or believe about a fact or a hypothesis and such categorization can be helpful for researchers while proving hypothetical assumption. Climate change conspiracy theories are popular in Twitter social media nowadays and it has a significant challenge for government and environmental organizations that are attempting to convince people to take actions against global warming, [8]. For example, an opinion holder may express his denial and believes about *Climate Change is a Hoax or not*, or *Climate change is caused due to population growth or not*. This is very much similar to detecting the stance of an author or speaker towards a discussion or debate about a specific topic.

Table 1: Climate Denier Twitter Account

Account Name	
wattsupwiththat	*can_climate_guy*
EcoSenseNow	*Cartoonsbyjosh*
ClimateRealists	*luisbaram*
SteveSGoddard	*sjc_pbs*
ClimateDepot	*co2science*
Carbongate	*Co2Coalition*
iowahawkblog	*GrizzlyGovFan*
FriendsOScience	*GrandSolarMin*
tan123	*TempGlobal*
CFACT	*ToryAardvark*
RupertDarwall	*IowaClimate*
BigJoeBastardi	*ClimateDepot*
JunkScience	*mattwridley*
BjornLomborg	

Source: BFTCD -Top 10 Climate Deniers

In the research work of [2], twitter accounts that are climate denier tend to be most likely conservative and follower of another climate denier account, links, blogs and websites. Refer table 1, for climate change denier accounts. Accounts who deny climate changes as a topic will never accept the argument: *climate change is caused due to population growth.* Many people believes that climate change is caused due to increase in human population and thereby increase carbon emission, deforestation etc. However, there are similar groups that deny the above facts (BFTCD -Top 10 Climate Deniers). It has also been exposed in various articles that climate denier accounts in Twitter are Bots.

In the research work of [3], group of Twitter users having similar interest and characteristics are identified in two approaches i.e., using Twitter follower network which is a *graph-based approach* and another one is *semantic approach* which is based on analyzing profile's meta-data such as bio, profile name and URL etc. Twitter users who are interested in propagating or receiving information on specific topics are referred to as Topical Group.

According to [6]; the presence of hash tags in twitter is one of the important features in describing a topic and by creating a strong hash tag predictor list and domain specific keywords, the twitter dataset can be classified. Apart from twitter accounts, the classification of denier and believer can be done using hash-tag mentioned in the user's tweets. Similar approach for detecting Spam and Not Spam using hash tag has been used in [7]. In the research work of [1], a list of definitive hash tags was used. Below hash tags for Believer and Denier are predefined and definitive hash tag related to climate change tweets. Researcher in [10], uses propagation algorithm to detect denier and believer hash tag.

Semi-Supervised Machine Learning algorithm are used to train models when we have both small number of labeled data and very large unlabeled data. A small set of data is used to classify using unsupervised algorithm or annotated automatically. Then the small existing labeled data is used to classify the rest unlabeled data. The research work in [9], implemented three semi-supervised multi-class classification algorithms for *climate change* related tweets and they are self training, semi-supervised SVM, and Multinomial Naive Bayes. In their work Multinomial Naive Bayes model along with unigram feature ultimately had the best performance.

III. DATASET

The extracted unlabeled twitter dataset is related to climate change and population growth tweets for the last 10 years. All the tweets contain the words "climate change" and "population". It comprises of meta data such as Twitter ID, Tweet Text, Bio-data, Account Screen Name, Location etc. The complete data extraction technique and processing of missing information are shown in [4]. The total number of tweets used in this research work is 168,303.

IV. PROPOSED METHODOLOGY

A. Annotation

The unlabeled dataset is annotated at two levels and the final annotation is set by combining the two approaches i.e., Hash Tag + Seed Phrase Annotation and Account Based Annotation.

1. Hash Tag + Seed Phrase Annotation

Table 2: Believer Hash-tags

Hash Tags	
#climatechangeisreal	*#climatechangeisfalse*
#actonclimate	*#climatechangenotreal*
#extinctionrebellion	*#climatechangehoax*
#climateemergency	*#globalwarminghoax*
#climateactionnow	*#tcot*
#FactsMatter	*#ccot*
#ScienceMatters	*#ClimateHoax*
#ScienceIsReal	*#Qanon*

Table 3: Denier Hash-tags

Hash Tags	
#tlot	*# hoax*
#pjnet	*#carbontax*
#rednationrising	*#scam*
#votered	
#libtard	
#libtards	
#YellowVests	
# fakenews	

Tweets normally comprises of one or more hash tags. These hash tags are extracted from tweet text as well as user's biodata information. There are also tweets that do not have any hash tags. For those hash tags in the predefined list (Refer table 2 and 3) are directly mapped as Denier or Believer. If a hash tag does not exist in the predefined list, then it is search in association with the seed phrase list (Refer table 4 and 5) within the entire dataset. The seed phrase is searched using wild character search *(e.g., %climate%change%due%to%population%).* If the search is found then the hash tag is decided as denier or believer based on the category of the seed phrase. If no match is found then the hash tag is assigned as Neutral. The final annotation of the tweet is based on the preference rule or score *(denier>believer>neutral)* defined for the entire hash tags present in a tweet.

Table 4: Believer Seed Phrase

Phrase Word
climate change due to population
global warming due to population
climate change is cause by population
global warming is cause by population
climate change is related to population
global warming is related to population
population increase cause climate change
population growth cause climate change
population aging a factor climate change
population aging a factor global warming

Table 5: Denier Seed Phrase

Phrase Word
climate change not due to population
global warming not due to population
climate change is not cause by population
global warming is not cause by population
climate change is not related to population
global warming is not related to population
population not a factor of climate change
population aging not a factor for climate change
population aging not a factor for global warming

The annotation function *f(#x, dsp, bsp)* can be defined as, if #L is the list of hash tags in a tweet *t* and #x is an individual hash tag belonging to *#L*. Then the hash tag *#x* can be associated with one or more believer seed phrases *bsp* or denier seed phrases *dsp* in the entire dataset. The association of seed phrase with *#x* is being search using wild character % for each term in the phrase using Microsoft TSQL procedure.

$$f(\#x, dsp, bsp) = \begin{cases} Denier, \sum count(\#x\ , dsp) > \sum count(\#x, bsp) \\ Believer, \sum count(\#x\ , dsp) < \sum count(\#x, bsp) \\ Neutral, \sum count(\#x\ , dsp) = \sum count(\#x\ , bsp) \end{cases} \quad (1)$$

In table 6, the first tweet id has three hash tags and their calculated hash tag classes. As the denier tweets are very less compare to believer tweets (Refer table 7 for hash tag annotation), preference is given to denier tag to classify whether a tweet belongs to believer or denier or neutral. Hence the above 4 tweets belong to denier class. The reason for giving preference to the denier class is due to the fact that climate change deniers are expected to deny the fact that *"climate change is cause due to population growth"*. If a tweet does not have any denier hash tag, then preference is given to believer tag and finally for neutral tag. If a tweet does not contain any hash tag, then it is considered as Empty Tag. Neutral Tag and Empty Tag are further annotated using the Account based annotation.

Table 6: Example of Tweets Hash Tag and its corresponding classes

ID	Hash Tag List	Hash Tag Class
100402	#savecalifornia, #votedemsout, #votered	Neutral, Neutral, Denier
100189	#tcot, #gop	Believer, Denier
100708	#tcot	Denier
100720	#americans, #berkley, #climate, #climatechangehoax	Neutral, Believer, Denier

Table 7: Hash Tag Final Annotation & its percentage

Class	Tweet Count	Percentage

Denier	455	0.27%
Empty_HashTag	125,770	74.73%
Neutral	18,808	11.18%
Believer	23,270	13.83%
Total	168,303	

2. Twitter Account Based Annotation

Account based classification are based on the predefined list of denier account as specified in Table 1. It is applied in addition to hash tag annotation. 23 Denier hub account were used for the annotation. If a tweet or user description/bio or user's screen name contains any of the listed account then the tweet is annotated as Denier. Once the account-based classification is completed, the next step is to derive the final annotation rules based on the Hash tag + Seed phrase annotation (DB_#) and Account Based annotation (DB_@). Denier hash tag or accounts are given higher preference than any other classes such as Believer or Neutral. Below rule are used for annotation. Believer tweets are represented by 0 and denier as 1. Neutral tweets are excluded while training the model. Table 8 shows the final output (DB) after applying the preference rule.

```
WHEN [DB_@]='Denier' THEN 1
WHEN [DB_#]='Denier' THEN 1
WHEN [DB_#]='Believer' THEN 0
WHEN [DB_#]='Neutral' THEN 'Neutral'
```

The final label i.e., **DB** is illustrated in table 8.

Table 8: Output Sample after applying final annotation rule

DB_#	DB_@	DB
Believer	Denier	1
Neutral	Denier	1
Denier	Denier	1
Believer		0
Neutral		Neutral
Denier		1

Table 9: Final annotated classes, tweet count and percentage

Class	Tweet Count	Percentage
Denier	919	0.55%
Empty_HashTag_ Account	125,443	74.53%
Neutral	18,749	11.14%
Believer	23,192	13.78%
Total	168,303	

In table 9, after account Based annotation step, it does not show much significant increase in the total number of tweets or percentage belonging to Denier and Believer class. This might be due to fact that the small number of accounts hub were used. We can further analyze the followers of those hub accounts to get more denier accounts and apply the above account-based rules. Twitter accounts that follow denier hub account are more likely to be a denier of *climate change* related topic as well as denier of the hypothesis that *climate change is due to population growth*.

B. Supervised Classification

1. Preprocessing

In machine learning, preprocessing of input data is an important step for effective and optimum error free learning. Word tokens are converted into numbers and it is the input for various supervised machine learning classification algorithms. Social media data are not always clean and formatted. It contains noises and junk values. Cleaning of such noises improves the accuracy and efficiency of the machine learning model. In this research paper the cleaning steps comprises of below steps.

a. Removing of Line Breaks
b. Case conversion to small letter
c. Removing of URLs
d. Removing numeric values
e. Punctuations and special characters
f. Removing twitter account
g. Removing Stop Words

2. Word Vectorizer

Supervised Machine learning algorithms operates on numerical data and feature spaces. Documents or tweet sentences are converted into rows (tweet text) and features columns (words or tokens) which are normally termed as vector representations before the actual machine learning algorithms are applied for training the dataset. This process of feature extraction is called as Vectorization. This paper focus on two vectorizers namely *Count Vectorizer* and *TF-IDF (Term Frequency & Inverse Document Frequency)* which are frequency-based embeddings.

In *Count Vectorizer* (One-Hot Encoding), a corpus *C* of having *D* documents i.e *{d1, d2, d3…dN)* and *n* is the number of unique tokens extracted from the corpus *C*, then the dictionary or the count vector matrix *M* formed by *n* tokens will have the size of *D* times *n*. Each row in the matrix *M* represents the frequency of tokens in document *D(i)*. The result of such matrix will be very large if the size of the dataset is large however it will contain the accurate count of words. A Count

Vectorizer primarily focuses on the occurrence of a word in a single document however in *TF-IDF*, it focuses both at the document level as well as the entire corpus. TF-IDF comprises of two terms i.e., *Term Frequency (TF)* and *Inverse Document Frequency (IDF)* [17]. TF measure how frequently a term occurs in a document and IDF measure how important a term is in the document. IDF measure the importance of document in the whole set of corpora. Document frequency of a term *t* is the occurrence of the term *t* in document. Formula for TF and IDF can be defined as:

$$TF(term\ t) = \frac{Count\ of\ term\ t\ appeared\ in\ a\ Document}{Total\ No.of\ terms\ in\ a\ Document} \quad (2)$$

$$IDF(term\ t) = \log_e(Total\ Number\ of\ Document\ /\ No.of\ Documents\ with\ term\ t) \quad (3)$$

3. Baseline Supervised Classification Models

In this paper, different classification models are applied on the annotated subset of the labeled dataset. The performance and computational efficiencies are analyzed for each model. The best performing model with respect to our available annotated dataset is identified and selected to predict remaining unlabeled data. After the whole dataset is labeled, it can be further analyzed and proof the assumed hypothesis. Best performing Supervised models such as *Logistic Regression, Multinomial Naïve Bayes, Linear Support Vector Machine (SVM) and Decision Tree* are used along with different set of parameters such as unigram, bigram, trigram features.

Logistic Regression is a machine learning algorithm for classification problem and it is based on discrete set of classes or categorical data. (Indra [11]). If the classification is only 2 then it is *Binary Logistic Regression* and if it is more than 3 or more class it is called as *Multinomial Logistic Regression*. It is primarily based on predictive analysis algorithm and the concept of probability. A cost function (sigmoid activation function) is used in Logistic regression and has a function value of limit from 0 to 1. The sigmoid function can be represented by $\sigma(z) = 1/(1+e^{-z})$, *where* $z = bias + w_1x_1+w_2x_2+\ldots+w_nx_n$ and w_i is the weight of x_i feature. Fig. 1, represents a sigmoid function graph. If the prediction function returns a value above or below the threshold value (Decision Boundary) then the class of the observation is determined. This algorithm is suitable for the research work as the classification problem is to predict Denier and Believer tweets related to *climate change and population*. For example, if the *predicted value >= 0.5* threshold value then the tweet will be classified as Denier (1) else Believer (0).

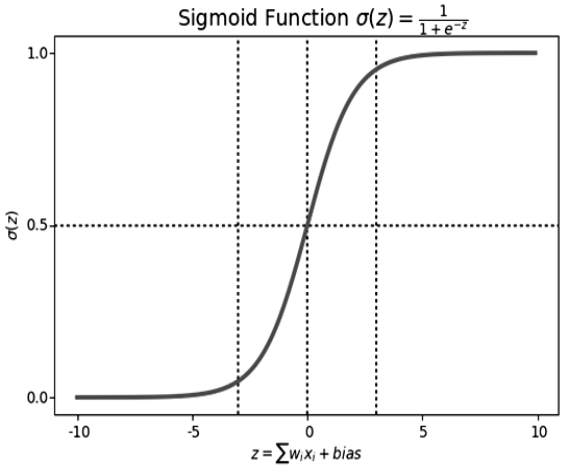

Figure 1: Sigmoid Function Graph

Multinomial Naïve Bayes (NB) algorithm is another supervised probabilistic learning method. Refer Manning et al. 2008. The probability of a document *d* being in class *c* can be computed as: $P(c|d) \propto P(c) \prod_{1\le k\le n_d} P(t_k|c)$, where $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class *c*. It measures how much evidence the term t_k contributes that *c* is the correct class. The prior probability of a document occurring in class *c* is denoted by P(c). The best class in NB classification is the most likely class C_{map} (Maximum a Posteriori class). i.e.,

$$C_{map} = \arg\max \hat{P}(c|d) = \arg\max \hat{P}(c) \prod_{1\le k\le n_d} \hat{P}(t_k|c),$$

where $\hat{P}$ represents the estimated probability. There can be floating point overflow as this equation has multiplication of many conditional probabilities. Applying logarithm will still hold the highest log probability score. i.e.

$$C_{map} = \arg max\ [\log \hat{P}(c) + \prod_{1\le k\le n_d} \log\ \hat{P}(t_k|c)] \quad (4)$$

The prior probability of class *c* can be computed as $\hat{P}(c) = \frac{N_c}{N}$, N_c is the total number of documents in the class *c* and *N* is the total number of documents. The conditional probability $\hat{P}(t|c)$ is computed as the relative of term *t* in documents belonging to class *c*. i.e.

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t \in V} T_{ct}} \quad (5)$$

Laplace smoothing is applied to remove the problem of sparseness or zero probability which normally occurs when the probability of a term is zero.

Linear Support Vector Machine, [5], is a vector space-based AI strategy and the essential target is to discover a decision boundary between two classes that is maximally far from any point in the training data. SVM can be extended to multiclass problems and non-linear models. The decision from the decision surface to the closest data point determines the margin of the classifier. Points located at the maximum margin are called as support vectors. It uses the stochastic gradient descent (SGD) method for optimizing the objective function.

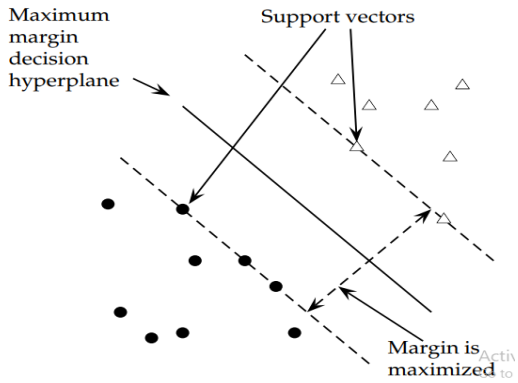

Figure 2: Support vectors and hyper-plain

A decision hyperplane can be defined by an intercept term b and a normal vector $\vec{w}$ which is perpendicular to the hyperplane. This vector is also referred to as weight vector. Refer fig. 2. All hyperplanes which are perpendicular to the normal vector will satisfy the linear equation $\vec{w}^T\vec{x} = -b$ where $\vec{x}$ represents all points on the hyperplane and $\vec{w}^T = \vec{w}/||\vec{w}||$. Suppose $D = \{(\vec{x}_i, y_i)\}$ is a set of training data points where each member is a point $\vec{x}_i$ and a class label y_i corresponding to it. Then the two classes in SVM classifier will be represented by +1 and -1 and the intercept term is explicitly represented as b. The linear classifier can be derived as $f(\vec{x}) = sign(\vec{w}^T\vec{x} + b)$.

Lastly, *Decision Tree Model* [12], is a distribution free model, which means that it is a non parametric method that does not depend on probability distribution of words or tokens. It is suitable for high dimensionality data and gives good accuracy. The internal nodes in the decision tree represents feature or attribute, its branches represent decision rules and the leaf nodes represents the outcome or result. Refer fig. 3. It has a root node at the top and the tree is partitioned recursively based on the attributes. The selection of the attribute is based on *Attribute Selection Measure (ASM).* Information Gain is one of the techniques. The initial step is to pick a target attribute or class attribute which will be our class i.e., Believer or Denier. Next step is to calculate the *Information Gain* IG of that *target attribute*. The formula is given by:

$$IG = -\frac{P}{P+N} log_2(\frac{P}{P+N}) - \frac{N}{P+N} log_2(\frac{N}{P+N}), \quad (6)$$

N and P represent count of each category within the attribute. For finding root node, Maximum Information gain for each category within an attribute will represent the Entropy of the attribute. Entropy will be calculated for all the remaining attributes. Once we have the Entropy then the Gain of the attribute is calculated. The maximum Gain of all attributes will be the root of the decision tree. Entropy of an attribute A is given by:

$$E(A) = \sum_{i=1}^{n} \frac{P_i+N_i}{P+N} I(P_iN_i), \quad Gain = IG - E(A) \quad (7)$$

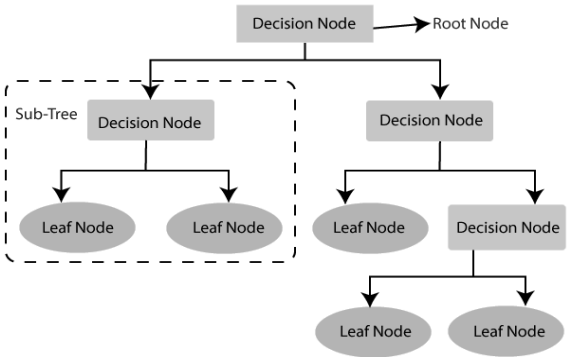

Figure 3: Decision Tree Structure

C. Proposed Algorithm

The dataset containing *climate change and population related tweets* are annotated using the above proposed automatic combine annotation method. Not all the dataset are annotated using this technique. This technique is a semi-supervised learning method where the remaining un-annotated data will be predicted using the annotated data. The proposed algorithm is stated as below:

a. Apply Hash Tag based annotation rule on the dataset to classify Believer, Denier & Neutral
b. Apply Twitter Account based annotation rule on the dataset to classify Denier.
c. Apply Seed word or phrase in combination with step *a* to classify Believer or Denier.

d. Apply final rules to classify Believer or Denier using step *a, b* and *c*.
e. Preprocessing of Twitter Text and removal of stop words.
f. Extract the subset of annotated tweets whose labels are Believer and Denier only
g. Create training and testing dataset in the ratio 80:20.
h. Create Count Vectorizer for unigram, bigram and trigram for tweets.
i. Create Term Frequency and Inverse Document Frequency vectorizer for unigram, bigram and trigram for tweets.
j. Initialize models Logistic Regression, Naïve Bayes, Linear SVM and Decision Tree.
k. For each model
a. Fit or train the model using Count Vectorizer and TF-IDF Vectorizer
b. Calculate Accuracy and F1 score for each model and vectorizers against test set
l. Pick the best model and parameter used.
m. Apply the trained model to predict remaining dataset that are not labeled as Believer and Denier.

Below table represents vectorizer notation using different n-gram combination.

Table 10: Vectorizer Notation and its parameters

Vectorizer Notation	Parameter
CountVectorizer_1	Default Unigram
CountVectorizer_2	Unigram + Bigram
CountVectorizer_3	Bigram + Trigram
TfidfVectorizer_1	Default Unigram
TfidfVectorizer_2	Unigram + Bigram
TfidfVectorizer_3	Bigram + Trigram

V. RESULT AND DISCUSSION

Table 11: Top 5 training outcomes of models.

Vectorizer Type	Model Name	Accuracy	Precision	Recall	F1-score
CountVectorizer_2	LinearSVM	***97.33***	82.35	38.04	52.04
CountVectorizer_1	LinearSVM	97.31	89.71	33.15	48.41
TfidfVectorizer_2	LinearSVM	97.18	94.44	27.72	42.86
CountVectorizer_2	Naive Bayes	96.16	49.63	36.41	42.01
CountVectorizer_3	LinearSVM	97.08	86.44	27.72	41.98

Table 12: Confusion Matrix of the highest accuracy and F1 score.

Confusion Matrix		Predicted	
		Believer (0)	Denier (1)
Actual	Believer (0)	4624	*15*
	Denier (1)	*114*	70

In table 9, the total number of annotated tweets (believer & denier label) is 24,111 and it is around 14.33% of the entire tweets collected (Train Set: 19,288 and Test Set: 4823). This subset of tweets is annotated with labels Believer (0) and Denier (1). The number of believers is higher than denier tweets. Neutral tweets are excluded as it will not make much sense in training the model. Empty Hash Tag Account tweets are those pure tweets that do not have any predefined hash tag and does not belong to any denier twitter account.

As per the result, the accuracy of the model is above 95% for each model. Linear SVM with Count Vectorizer of unigram and bigram features combination score the highest accuracy and TFIDF vectorizer for unigram and bigram is next to it. Accuracy is the measure of all the correctly identified classes. The top 5 outcomes of the experiment are shown in table 11. The complete scores are shown in fig. 4 for each model. In Table 12, False Positive and False Negative (i.e., False Believer and False Denier) are high and thereby the cost is high. This score needs to be reduced. All though the overall accuracy of SVM (Countvectorizer2) i.e., 97.33% is high enough but F1 Score (Harmonic mean of Precision and Recall) i.e., 52.04% is not up to the mark. As the dataset is an imbalance class dataset, the performance of both precision and recall are not high. Recall is low when False Negative and False Positive cost are high. If both precision and recall are high then the F1 score could have been higher. Refer [13], [14].

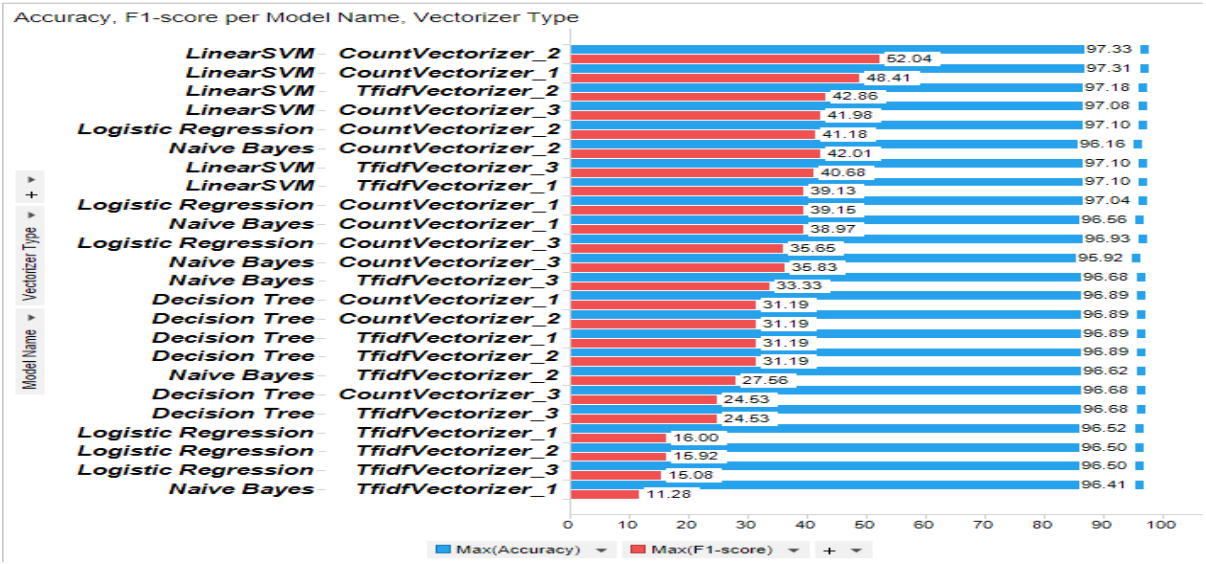

Figure 4: Graphical representation of Accuracy and F1-score for each model with different hyper parameter.

VI. CONCLUSION

The annotation process used in this experiment needs to further tune for finding denier tweets. Especially Twitter Account Based Denier technique of annotation can be improvised by further identifying more followers to a Denier account. Another area of improvement can be redefining the preference rule for hash tag-based annotation. With this the imbalance class in the training dataset can be improved. Considering neutral tweets as denier could be the area of focus for future experiment or it can be further classified as "Denier" or "Believer" using unsupervised techniques. However, the accuracy 97.33% of the selected supervised model i.e., Linear SVM with count vectorizer of Unigram and Bigram outperform other baseline models used in this research paper.

REFERENCE

[1] Andrew Graves, (2020), "Classifying Climate Change Tweets", Utilizing NLP and classification techniques to categorize tweets as climate change believer or denier tweets, Available: https://towardsdatascience.com /classifying-climate-change-tweets-8245450a5e96

[2] Jeremy McGibbon, (2020), "Connections among climate deniers on Twitter", Available: https://courses.cs.washington.edu/courses/cse544/18wi/project/examples-successful-projects/mcgibbon.pdf

[3] Bhattacharya, P., S. Ghosh, J. Kulshrestha, M. Mondal, M. Bilal Zafar, N. Ganguly, and K. P. Gummadi, *"Deep Twitter diving: exploring topical groups in microblogs at scale"*, In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. Baltimore, MD: ACM, pp. 197-210, 2014.

[4] Chongtham Rajen Singh, R. Gobinath, *"Identify missing countries using GEEBLL iterative method for analyzing tweets opinion"*, Materials Today: Proceedings, ISSN 2214-7853, 2020.

[5] C.D. Manning, P. Raghavan and H. Schuetze, *"Introduction to Information Retrieval"*, Cambridge University Press, pp. 234-265, 2008.

[6] V. Gupta and R. Hewett, *"Harnessing the power of hashtags in tweet analytics," 2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, USA, pp. 2390-2395, 2017.

[7] Surendra Sedhai and Aixin Sun. *"An analysis of 14 Million tweets on hashtag-oriented spamming"*. J. Assoc. Inf. Sci. Technol. 68, 7 (July 2017), 1638–1651, 2017.

[8] Douglas, Karen & Sutton, Robbie. *"Climate change: Why the conspiracy theories are dangerous"*, Bulletin of the Atomic Scientists. 71. Pp 98-106, 2015.

[9] Kabaghe, C. (2019). Classifying Tweets Based on Climate Change Stance Natural Language Processing.

[10] Tyagi, Aman & Babcock, Matthew & Carley, Kathleen & Sicker, Douglas. "*Polarizing Tweets on Climate Change*", 10.1007/978-3-030-61255-9_11, 2020.
[11] Indra, S. & Wikarsa, Liza & Turang, Rinaldo. *"Using logistic regression method to classify tweets into the selected topics"*, 385-390, 2016.
[12] Nilosey, Shivam & Pipliya, Abhishek & Malviya, Vijay, *"Real-Time Classification of Twitter Data Using Decision Tree Technique"*, 10.1007/978-981-15-2071-6_14, 2020.
[13] Goutte, Cyril & Gaussier, Eric, *"A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation",* Lecture Notes in Computer Science. 3408. 345-359. 10.1007/978-3-540-31865-1_25, 2005
[14] Sokolova, Marina & Japkowicz, Nathalie & Szpakowicz, Stan, *"Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation",* AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science. Vol. 4304. 1015-1021. 10.1007/11941439_114, 2006.
[15] BFTCD, Before the Flood: Top 10 Climate Deniers, Retrieved From https://www.beforetheflood.com/explore/the-deniers/top-10-climate-deniers/
[16] TOPRO, The Overpopulation Project, Retrieved From https://overpopulation-project.com/population-growth-is-a-threat-to-the-worlds-climate/
[17] Robertson, S. (2004), "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of Documentation, Vol. 60 No. 5, pp. 503-520. https://doi.org/10.1108/00220410410560582