


## Article

# Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches

Nabila Mohamad Sham<sup>1</sup> and Azlinah Mohamed<sup>2,\*</sup> 

<sup>1</sup> Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam 40450, Malaysia; nabila.mohamadsham@gmail.com

<sup>2</sup> Institute for Big Data Analytics and Artificial Intelligence, Universiti Teknologi MARA (UiTM), Shah Alam 40450, Malaysia

\* Correspondence: azlinah@uitm.edu.my

**Abstract:** The emissions of greenhouse gases, such as carbon dioxide, into the biosphere have the consequence of warming up the planet, hence the existence of climate change. Sentiment analysis has been a popular subject and there has been a plethora of research conducted in this area in recent decades, typically on social media platforms such as Twitter, due to the proliferation of data generated today during discussions on climate change. However, there is not much research on the performances of different sentiment analysis approaches using lexicon, machine learning and hybrid methods, particularly within this domain-specific sentiment. This study aims to find the most effective sentiment analysis approach for climate change tweets and related domains by performing a comparative evaluation of various sentiment analysis approaches. In this context, seven lexicon-based approaches were used, namely SentiWordNet, TextBlob, VADER, SentiStrength, Hu and Liu, MPQA, and KWWSI. Meanwhile, three machine learning classifiers were used, namely Support Vector Machine, Naïve Bayes, and Logistic Regression, by using two feature extraction techniques, which were Bag-of-Words and TF-IDF. Next, the hybridization between lexicon-based and machine learning-based approaches was performed. The results indicate that the hybrid method outperformed the other two approaches, with hybrid TextBlob and Logistic Regression achieving an F1-score of 75.3%; thus, this has been chosen as the most effective approach. This study also found that lemmatization improved the accuracy of machine learning and hybrid approaches by 1.6%. Meanwhile, the TF-IDF feature extraction technique was slightly better than BoW by increasing the accuracy of the Logistic Regression classifier by 0.6%. However, TF-IDF and BoW had an identical effect on SVM and NB. Future works will include investigating the suitability of deep learning approaches toward this domain-specific sentiment on social media platforms.



**Citation:** Mohamad Sham, N.; Mohamed, A. Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. *Sustainability* **2022**, *14*, 4723. <https://doi.org/10.3390/su14084723>

Academic Editors: Ayyoob Sharifi, Baojie He, Chi Feng and Jun Yang

Received: 20 March 2022

Accepted: 11 April 2022

Published: 14 April 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** climate change; sentiment analysis; lexicon; machine learning; social media

## 1. Introduction

One of the most pressing issues of the 21st century is the warming up of the Earth caused by the emissions of greenhouse gases, such as carbon dioxide, into the atmosphere. This period started as the industrial revolution took place, and the consequences it has brought upon us are detrimental, especially to the natural environment and climate change. According to the World Meteorological Organization [1], 2020 was one of the warmest years ever recorded in human history, despite the reduction of greenhouse gas emissions due to COVID-19 measures. Furthermore, the year 2020 recorded high occurrences of extreme events, such as hurricanes, extreme heatwaves, severe droughts, and wildfires, which have led to population displacement and even deaths [1]. It is therefore imperative that our minds should work collectively in order to mitigate against further damage related to climate change and this can be achieved through investigating the sentiment of the population and policy changes.

Sentiment analysis, which is a branch of Natural Language Processing (NLP), is a field that focuses on discovering techniques to decipher the sentiments hidden in text comments from reviews or opinions posted online [2]. The main objective of sentiment analysis is to study and analyze the perception and opinion or viewpoint with respect to entities in a discussion. Among all of the possible mediums of discussion, analysis of the natural language by analyzing the sentiments contained within a text remains the most popular application in the sentiment analysis field [3]. Furthermore, by investigating the tone of words within a text through sentiment analysis, it can shed some light on people's reactions towards a particular topic that could either be negative, neutral, or positive, thus this insight can help the decision-makers to understand public behavior and how to tackle the issues at hand [4]. Sentiment analysis has been successfully applied to study different domains, for instance news headlines [5], movie reviews [6], presidential elections [7], and GST [8].

Today, due to the rapidly growing access to the Internet all over the world, there is a blooming of social media platforms which has attracted a large number of users to interact with each other and to publicly offer their thoughts or opinions on various things. This activity has generated a huge amount of structured and unstructured data which can be further analyzed to generate valuable insights. The popularity of social media platforms is increasing every year as the diffusion of knowledge is faster compared to traditional news outlets. One of the social media platforms that has gained tremendous popularity over the years is Twitter, and it is considered a gold mine for user-generated opinionated data. Moreover, these opinions or sentiments expressed by users have been proven to have an impact on society at large [9]. People from diverse backgrounds, such as NGOs, activists, celebrities, politicians, and even the general public, have expressed their stances and opinions regarding climate change issues on Twitter. The most popular attributes of climate change discussions include whether it is real or not and what are the possible mitigation strategies to combat the impacts of climate change [4]. Therefore, by utilizing the vast opinions available on this social media platform, researchers can investigate the evolution of sentiment regarding the social topics being discussed, such as climate change.

Sentiment analysis consists of two basic approaches, which are lexicon-based approaches and machine learning approaches. Lexicon-based approaches makes use of a pre-built dictionary containing words that have been tagged as either positive, negative, or neutral, and these words can also have sentiment intensity associated with them. Meanwhile, machine learning approaches involve statistical methods that employ feature extraction methods, whereby every word or multiword will be vectorized to be used as independent features during training [10]. Due to the variation of the words present in the texts as the result of informal language being used on the Twitter social media platform, these words may not be present in the lexicon dictionaries, thus the detection of the sentiment polarity will be ineffective and inefficient [11]. Therefore, some researchers have opted for other alternative methods, such as machine learning techniques that can be trained and later used for the predictions of sentiment polarity.

Lexicon-based approaches have the advantage of lacking the requirement for large training corpora, therefore when a sizeable training corpus could not be obtained, the researchers will opt for a lexicon-based approach [12]. Although lexicon-based approaches have the strength of being computationally efficient and scalable for determining the sentiment, this strength will be undermined when linguistic rules are applied [12]. One of the major drawbacks of lexicon-based approaches is the requirement for a large number of linguistic resources to be inspected in order to generate important words and their sentiment polarity within a domain as these words can influence the coverage and quality of the sentiment dictionaries [13]. Therefore, the main disadvantage of lexicon-based approaches is that performance will vary considerably across different domains [14]. Therefore, a lexicon that has achieved high accuracy in one domain may not perform well in another domain.

Machine learning approaches, however, depend on the domain that it has been trained on and it has been shown that one technique will not guarantee an optimal level perfor-

mance when used on another domain [15]. This would mean that if the classifiers have been trained on one specific corpus, they have to be retrained again if they are to be used on another domain in order to produce the same accuracy [16]. Furthermore, machine learning approaches have been shown to outperform lexicon-based approaches in most cases; however, they require a large training corpus that has been pre-labeled with sentiment which can be time-consuming but without it accurate results would not be obtained [17]. Several machine learning approaches have certain advantages based on the characteristics of the texts, for example Naïve Bayes and Support Vector Machine are more effective for short and full-length reviews, respectively [18]. According to Mahmood et al. [19], lexicon-based approaches are more robust and accurate when applied to various domains compared to machine learning approaches.

This study will perform comparative evaluations of various sentiment analysis techniques on climate change tweets, as well as applying a hybrid approach which combines both the lexicon-based and machine learning approaches to exploit the strengths of each method, thus enhancing the performance obtained. The combination will be developed by firstly obtaining the semantic orientations of the tweets given by the lexicon-based approach, and these outputs will be used as training data in the machine learning classifiers. The lexicon-based approaches include SentiWordNet, SentiStrength, MPQA, Hu and Liu, WKWSCI, VADER, and TextBlob. The reasons for choosing these lexicons include that they are frequently used lexicons in the sentiment analysis field, specifically on the Twitter domain. Furthermore, WKWSCI, Hu and Liu and MPQA lexicons have shown promising results with accuracies of above 80% in the customer reviews domain, but there is a lack of research that has utilized them on Twitter data. The performance of these lexicons will be compared against machine learning classifiers, such as Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR). Finally, the hybridization between the lexicons and machine learning classifiers will be performed to improve the overall performance of the sentiment classification. The research questions for this study are outlined below to answer the hypothesis, that climate change sentiment in social media will have a more significant impact on hybrid sentiment analysis approaches than on traditional sentiment analysis approaches, and also which approaches provide better accuracy for this specific domain.

#### *Research Questions*

- RQ1: What are the effects of lemmatization on the performance of sentiment analysis methods?
- RQ2: What is the influence of feature extraction techniques on the performance of machine learning-based approaches?
- RQ3: How is the performance comparison of various sentiment analysis approaches, which are lexicon, machine learning and hybrid methods, for classification of climate change tweets?

This study is divided into four sections: related work, methodology, results and discussions, and conclusions. Related work explores the literature review related to sentiment analysis. Methodology outlines the implementation steps of various sentiment analysis approaches adopted in this study. The results and discussion section provides the results obtained through the implementation of the sentiment analysis classification models, including the different techniques used. The summary of the discussions is provided in the conclusions of each section.

## **2. Related Work**

This chapter comprises of several topics, namely the understanding of Natural Language Processing (NLP) and sentiment analysis on Twitter, including its application. Next, it explores various sentiment analysis approaches comprising of lexicon-based and machine learning-based techniques. A review on data preprocessing and feature extraction techniques will follow afterwards.

### 2.1. Natural Language Processing Overview

In terms of big data, 95% are in the form of unstructured texts originating from web pages, email, etc., thus in order to exploit the advantages of this vast amount of data effectively, new scientific tools and methods are required [20]. NLP techniques have made this endeavor a reality due to its capabilities to harvest textual heavy data and can be considered as a form of artificial intelligence, as it makes use of computational algorithms, such as machine learning and deep learning, to learn, understand and ultimately generate semantically sound human language [20]. NLP has proven it can achieve comparable results that have been obtained through traditional qualitative text analysis techniques, such as the study by Guetterman et al. [21], which stated that NLP managed to discover major themes provided through traditional text analysis; however, it also failed at detecting nuances and remains inferior in terms of providing details in context. NLP is a multi-disciplinary field that has been adopted in healthcare, clinical and social media, among many other areas. Although there has been substantial growth in the application of deep learning methods into the NLP field, it is not widely used in clinical practice as the interpretability rate of these methods are still low, unlike the traditional machine learning methods [22]. Lexicon-based approaches can be categorized as ones of pure NLP techniques and these approaches have gained popularity in the sentiment analysis field, especially during 2016–2017. Meanwhile, the recent advances in the sentiment analysis field have seen the emergence of new techniques that are shifted from pure NLP to conventional machine learning classifiers, as well as deep learning-based models that have been shown to outperformed other methods by achieving state-of-the-art performances [23].

### 2.2. Sentiment Analysis on Twitter

The proliferation of data requires us to automatically analyze them in order to generate insights. Sentiment analysis deals with techniques that automatically classify human opinions embedded in text, which include sentiment and emotions that have been expressed by the people on various subjects, such as products reviews or opinions on energy sources [24]. Sentiment analysis can include problems such as three-class sentiment categorization tasks into either positive, negative, or neutral. Other studies focus on finding the strengths of the sentiment class categories, such as extremely positive, positive, neutral, negative and extremely negative [25], and emotions, such as “angry”, “sadness”, “fear” and “joy”, expressed. Sentiment analysis can be performed at three distinct levels, i.e., aspect, sentence, or document level, and NLP techniques are often employed to analyze those texts automatically. Document level aims to identify the sentiment of the whole document, meanwhile sentence level is more fine-grained and aims to analyze the sentiment of an individual sentence. Another level of sentiment analysis is the aspect level, which aims to identify the sentiment of an attribute that are usually being discussed together during a product review process [23].

Social media platforms that have attracted millions of users worldwide provide the opportunity for people to share their thoughts on certain subjects of interest. As of 2021, a tweet can contain up to 280 characters, therefore users need to summarize their opinions concisely within that limit. The difficulty in analyzing Twitter text includes noisiness in the data, such as spelling mistakes, grammatical errors, and the use of emoticons or slang, which change over time [25]. The sentiments derived through Twitter sentiment analysis have many applications, such as finding correlations between sentiments expressed with polling data during the 2016 presidential election [7], reviews on the tourism industry [26], pilgrimage season [27], and the emotions of people during the COVID-19 pandemic [28]. Sentiment analysis, which is also known as opinion mining, combines several NLP techniques, such as tokenization, stemming or lemmatization and Bag-of-Words, with machine learning models to calculate the sentiment scores of the terms within a particular text [29]. These sentiment scores in return can be used to determine the sentiment polarity of the whole text that can either be positive, neutral, or negative.

### 2.3. Types of Sentiment Analysis Approaches

There are three types of sentiment analysis approaches, namely unsupervised lexicon-based, supervised machine learning-based, and hybrid approaches. In supervised approaches, the training of the classifiers is based on the labelled dataset.

#### 2.3.1. Lexicon-Based Approaches

One of the main approaches in sentiment analysis is the usage of sentiment lexicons for polarity classification. A sentiment lexicon consists of words or phrases that have been tagged with scores and these scores will be used to calculate the overall polarity or collective sentiment orientation of the sentiment task. Sentiment lexicons are generated either by manually coding the sentiment of the associated words, investigating the semantic relations between a set of seed words from an existing sentiment lexicon, adapting the sentiment lexicon from one specific domain to another through transfer learning, or by using probabilistic approaches, such as machine learning classifiers to identify sentiment-bearing words [18]. Some of the pre-existing and frequently utilized sentiment lexicons are SentiWordNet and MPQA [30].

SentiWordNet has been built based on the English lexical dictionary WordNet, whereby words are grouped into synsets (synonym sets) and each synset has been assigned with three polarity scores, representing positive, negative, and neutral classes [25]. Agarwal et al. [8] used SentiWordNet alongside newspaper articles for the creation of domain-specific lexicon (financial) and discovered that some words that are tagged as neutral have certain polarity in other domains, thus confirming the notion that the use of lexicon is domain-dependent [31]. SentiStrength is proposed by Thelwall et al. [32], in which the polarity strength values range from 1 to 5 and makes use of emoticons, negations and boosting words for polarity detection. SentiStrength has been regarded as one of the best performing general-purpose lexicons by Zimbra et al. [14] as it can achieve an average classification accuracy of 66% when tested against Twitter datasets from various domains, such as pharmaceuticals and telecommunications. Multi-Perspective Question-Answering Subjectivity Lexicon (MPQA) is an aggregated lexicon originating from a variety of sources that has been constructed manually as well as automatically by Riloff and Wiebe [33]. MPQA has attained an accuracy as high as 73.20% when tested on news headline corpus (SemEval-2007 Task 14 “Effective Text”) to investigate the effectiveness of the lexicon on other domains different to product reviews.

Hu and Liu Opinion Lexicon is an automatically generated lexicon developed based on customer reviews compiled from multiple domains by using machine learning techniques [18]. Meanwhile, WKWSCI Sentiment Lexicon is a general-purpose sentiment lexicon proposed by Khoo and Johnkhan [18] and is developed by manually coding each word as positive, negative, or neutral, and its associated polarity strength ranging from 1 to 3 (slightly positive, positive, and very positive). The performances of multiple lexicons, namely WKWSCI, Hu and Liu, MPQA and SentiWordNet, have been compared in Khoo and Johnkhan [18] to investigate whether domain-specific lexicon can achieve reasonable results when applied on another domain, and the results obtained show that Hu and Liu outperformed the other lexicons in term of accuracy, with SentiWordNet performing the worse.

Valence Aware Dictionary for sEntiment Reasoning (VADER) is a parsimonious rule-based and human-validated sentiment analysis tool that has been constructed from generalizable and valence-based lexicon to be used for sentiment categorization of text on social media platforms, specifically Twitter. VADER takes into account grammatical and syntactical rules during its sentiment categorization tasks. VADER has been shown to achieve state-of-the-art accuracy, which is above 90%, when compared with other lexicons and outperformed machine learning methods [6,34]. TextBlob lexicon analyzes each word in the text and returns a tuple containing (polarity, subjectivity) in which the polarity ranges between  $-1$  and  $1$ . TextBlob underperformed when compared to both SentiWordNet and VADER due to its coverage problem as it tries to assign many messages as neutral [6].

This study will investigate the performance of TextBlob, VADER, SentiStrength, SentiWordNet, WKWSCl, Hu and Liu, and MPQA lexicons on classifying sentiment of tweets about climate change, in which none of the above methods are built for this domain. The existing approaches are built with the objectives of being general-purpose lexicons by containing sentiment-bearing words that are general in nature. These lexicons will be tested on climate change datasets, and the study examine their performance as is and when they are hybridized with machine learning classifiers.

### 2.3.2. Machine Learning-Based Approaches

Machine learning-based approaches consist of three categories, which are unsupervised learning that analyzes the words within the documents, and common words are then grouped together to form multiple clusters that have been specified by the researchers. Meanwhile, semi-supervised learning utilizes both unlabeled and limited numbers of labelled data during the training process of the classifier, whereby the unlabeled data are used to extract important features and the labelled data are used to fine-tune the output of the classifier [35]. This study will investigate the third category, which is supervised learning.

There have been many studies to improve sentiment analysis techniques and one of the approaches uses supervised machine learning classifiers. This technique is associated with the use of labelled data as features for training phases and learning algorithms to produce the expected output [36]. The sentiment polarity of the unlabeled data can then be predicted subsequently by using the trained classifier. One of the major drawbacks of this approach is the requirement for a large number of labelled data to be used during the training process, which can be inefficient if the data have to be labelled manually by humans, although it has been shown through a lot of studies that supervised learning outperformed both semi-supervised and unsupervised learning in many aspects [35]. A sample of 3102 tweets related to COVID-19 were tested using ML with a TF-IDF feature extraction technique and the maximum achieved accuracy for SVM was 80%, NB with 78% and LR with 76% [37]. A Kaggle dataset on airline reviews was used with ML classifiers and obtained an accuracy of 83.31% for SVM, 81.81% for LR and 73% for NB [38].

Due to the requirement of large-labelled datasets for the training of machine learning classifiers, several studies have opted for hybrid methods in sentiment analysis by using the polarity supplemented by lexicon-based approaches and used it for training the ML. Lalji and Deshmukh [39] noted the high precision and low recall of the lexicon methods and performed a hybridization between MPQA and SVM to improve the performance. Rajeswari et al. [40] concluded that the use of lexicon in hybrid approaches resolves the neutral class classification problem and the overall improvement in the accuracy when tested against product, Twitter, and movie datasets with NB, LR, SVM and DT classifiers. Furthermore, the authors also highlighted the importance of the feature extraction process and found that TF-IDF yields the best accuracy.

## 2.4. Data Preparation Techniques in Sentiment Analysis

The data preparation techniques include data pre-processing for both lexicon-based and machine learning-based approaches. Meanwhile, the feature extraction techniques are only applicable for machine learning-based approaches.

### 2.4.1. Data Preprocessing Techniques

The limited number of characters available and the irregular structure of text on Twitter present challenges to sentiment analysis. Furthermore, social media texts can consist of noises and uninformative terms, such as HTML tags, that can increase the dimensionality of the training data for machine learning classifiers [41]. One of the crucial steps in sentiment analysis is data preprocessing, which comprises of several methods and selecting the appropriate methods can enhance the system accuracy. Krouska et al. [42] performed a comparative analysis between several pre-processing methods and concluded that the feature selection technique increases accuracy rates significantly, while the overall

pre-processing techniques implemented are not biased to the domain of the datasets used. According to Angiani et al. [41], the most frequently used data pre-processing methods for tweets are tokenization, stemming, removal of punctuation marks and stop words. Pradha et al. [43] compared several preprocessing techniques, such as stemming, lemmatization and spelling correction. The authors found that lemmatization produced the lowest rate of sentiment dissimilarity between processed and unprocessed texts, followed by spelling correction and stemming. Meanwhile, Zhao and Gui [44] suggested the removal of stop words due to some of them containing different sentiment polarity.

#### 2.4.2. Feature Extraction Techniques

Two of the data preparation techniques in machine learning-based approaches for sentiment analysis are feature selection, which is a process of selecting the relevant features to be used during the training phase of the machine learning classifiers, and feature extraction, which is a construction of new features from the existing features. The latter will reduce the dimensionality of the features to a lower dimension, as higher feature dimensions are difficult for the machine learning classifiers to train, thus feature extraction techniques are one of the determinant factors that will influence the performance of the sentiment classification models [45]. The simplest method for feature extraction is Bag-of-Words (BoW), where each unique word and the frequency of its occurrence are taken together to generate a vocabulary [46]. Other feature extraction methods include TF-IDF, N-grams and Word2Vec, which build the vector representation of the words within the documents; however, the similarity between all of these methods is that these techniques assume the independent nature of each word, therefore they lack information regarding the syntactic and semantic relationships between those words [45]. Rustam et al. [11] investigated text classification methods by using machine learning on US airline review datasets from Twitter. The three different feature extraction techniques that were investigated were TF, which utilizes only the frequency of occurrences of each word, TF-IDF and Word2Vec. The findings from this study indicate that TF-IDF is the most appropriate feature extraction technique for the sentiment classification of Twitter data.

### 3. Methodology

This study aims to find the most effective sentiment analysis approach for the classification of climate change tweets and related domains. The tweets were first pre-processed according to two different paths, namely with and without lemmatization, before being input into the sentiment classification model. Then, the performances of the machine learning algorithms were further analyzed according to two different feature extraction techniques, namely Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). This proposed methodology was used albeit to the normal procedures in performing sentiment analysis. The flow of the implementation of the suggested approaches is as shown in Figure 1.

The implementation of the proposed methodology will be developed by using Python programming language, version 3.8.3, for the data preprocessing phase and lexicon-based approaches. Meanwhile, the training and evaluation of classifiers for both machine learning-based and hybrid approaches will be conducted by using the Orange data mining tool, version 3.28.0. Furthermore, SentiStrength version 2.3.7110.19972, which is copyrighted by Mike Thelwall [47] from the University of Wolverhampton, will be utilized to obtain the sentiment polarity of the tweets by using SentiStrength lexicon, and Hydrator-0.0.13 will be used to hydrate tweets IDs back into the original texts.

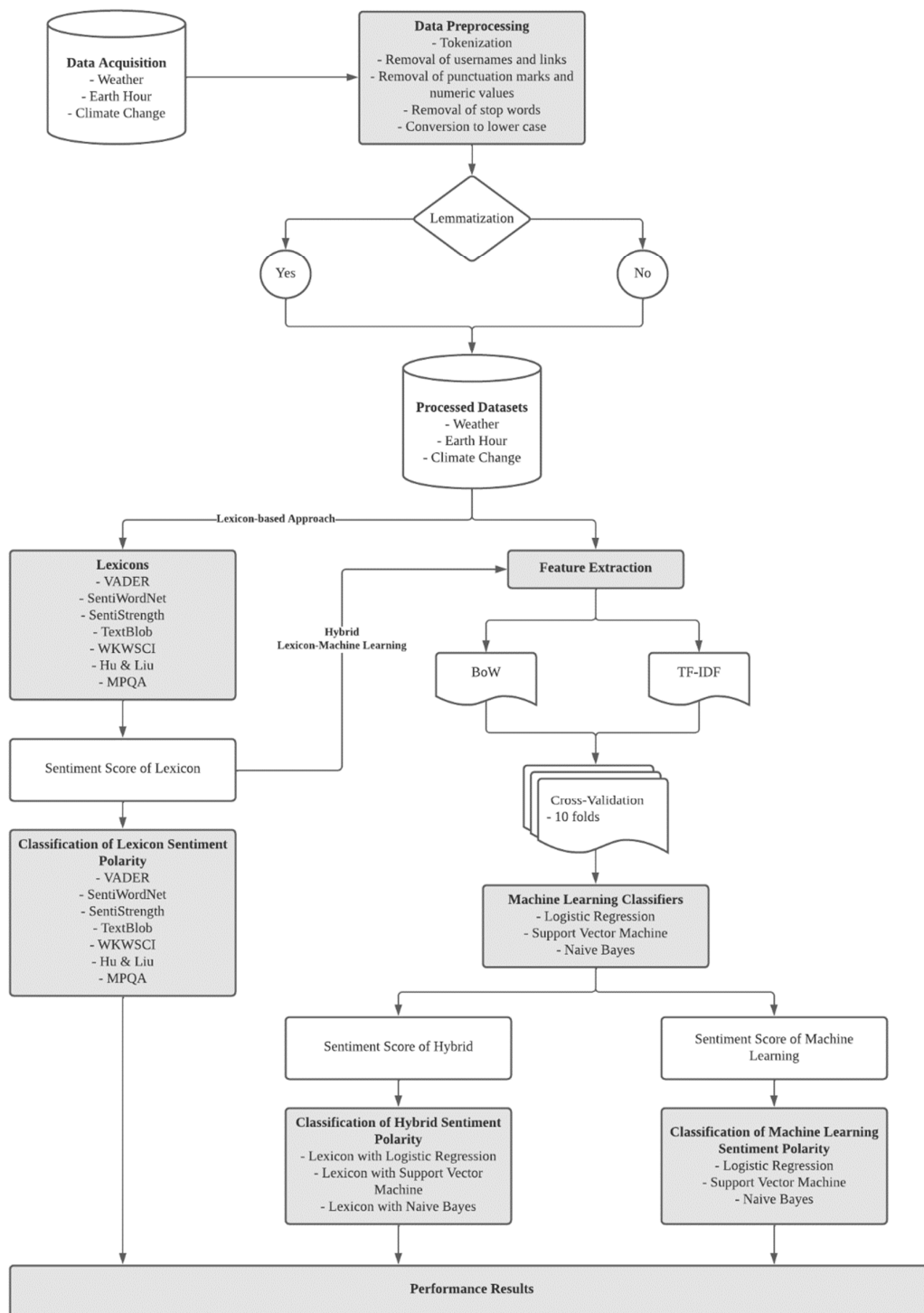


Figure 1. Climate change classification model design framework.

### 3.1. Data Understanding

Three Twitter datasets have been acquired to perform the comparative evaluation of different sentiment analysis approaches in this study. The data originated from tweets that have been scraped online through the Twitter social media platform and annotated with sentiment polarity. Each dataset consists of a different number of tweets within each sentiment class, namely negative, neutral, and positive.

The first dataset is “Weather Sentiment”, obtained from the data.world website [48], which contains 1000 tweets related to weather. The tweets were collected in 2013 and



20 contributors were asked to grade each tweet on its sentiment, which includes negative, neutral (author is just sharing information), positive and “I can’t tell”. For this study, only the tweets with negative, neutral, and positive sentiment polarity have been chosen for further analysis, resulting on 763 tweets in total.

The second dataset is “Earth Hour 2015 Corpus” obtained from the DecarboNet [49] database. This corpus was originally comprised of 600 tweets related to the Earth Hour 2015 event which took place on 28 March 2015, and it was used in the study by Maynard and Bontcheva [50] to evaluate the performance of different lexicons on analyzing environmental tweets. Each tweet was triple annotated through crowdsourcing. The downloaded corpus was in the form of dehydrated json file with 371 tweets and contains only the tweet IDs and its sentiment polarity of positive, neutral, or negative. Since the contents of the tweets were not provided, a crawler (Hydrator-0.0.13) was used to extract them from Twitter based on the tweet IDs, and this procedure returned only 252 tweets which indicates that some of them have probably been deleted by the authors of the tweets.

The third dataset is about climate change sentiment on Twitter obtained from the Kaggle website [51]. The tweets span from 1 January 2020 to 24 December 2020 and the dataset encompasses of 396 tweets in total. The tweets were queried by using the keywords “climate change” and “global warming”, with retweets being omitted from the tweets extracted. The tweets had been annotated with sentiment polarity by using TextBlob lexicon. The summary of the size of the datasets obtained is shown in Table 1. In general, the tweets that have been annotated to be neutral dominate over positive and negative tweets.

**Table 1.** Count of tweets within the three datasets discretized by the sentiment classes.

Dataset	Positive	Neutral	Negative	Total
Weather Sentiment	231	261	271	763
Earth Hour 2015 Corpus	64	162	26	252
Climate Change Sentiment	190	126	80	396
Total	485	549	377	1411

### 3.2. Data Preprocessing

Preprocessing of textual data is crucial because this step can influence the accuracy of the sentiment classification. Several data preprocessing techniques have been implemented on the acquired datasets to prepare them for further analysis, which includes tokenization, removal of usernames, links, punctuation marks, numeric values and stop words. Furthermore, every character will also be converted to lower case and lemmatization will be performed afterwards.

Tokenization is the process of splitting the text or string into tokens which can be made up of individual words, phrases, or paragraphs, depending on the level of tokenization performed. In this study, the level of tokenization chosen is the word level, which splits the text based on white spaces by using the TweetTokenizer function from NLTK.

Usernames of Twitter accounts which begin with the “@” symbol, any URLs (e.g., http), “RT” signs which indicate the tweets have been retweeted by other users, punctuation marks such as “#” and numeric values embedded within the tweets do not contribute to sentiment analysis and they can add noise during the training of machine learning classifiers. Thus, they should be removed promptly. According to Elbagir and Yang [52] hashtags (#) often contain useful information regarding the subject of the tweet (e.g., #awareness), therefore only its symbol will be removed from this study.

The characters will be converted to lower case for uniformity before removing the stop words. In this study, stop words to be removed are “this”, “is”, “are”, “earth”, “hour”, “weather”, “climate” and “change”. These are the words that were used as the keywords during the scraping process of the tweets. Stop words do not carry much useful information in them; therefore, they will be removed from further analysis [53].

Lemmatization involves the usage of a dictionary or lookup table and the context of the terms, to return them to their dictionary forms. Lemmatization will also match the synonyms of the terms, for example “cars” will return “car” as well as “automobile”, and this process happens by analyzing the morphological properties of the terms, i.e., whether they are used as verbs or nouns in the text [54]. Lemmatization will reduce the number of indexes used, thus decreasing the size or complexity of the features during the training of machine learning classifiers. Due to this strength of lemmatization, this study will also investigate the influence of lemmatization on the performance of the sentiment analysis approaches.

### 3.3. Sentiment Lexicon Evaluated

The sentiment lexicons that will be evaluated include SentiWordNet, Textlob, VADER, SentiStrength, Hu and Liu Opinion Lexicon, MPQA Subjectivity Lexicon, and WKWSCl. The implementation of lexicons, typically SentiWordNet, TextBlob, and VADER, will be conducted by using Python. Meanwhile, Hu and Liu, MPQA and WKWSCl lexicons will be implemented by using the Orange data mining tool.

#### 3.3.1. SentiWordNet

The tweets are first parsed individually to be tokenized into independent words or lexical terms. In the Part of Speech (POS)-tagging, each lexical term will then be either tagged as nouns, verbs, adjectives, or adverbs depending on its context and meaning. By analyzing the POS-tagged sentence, the structure of the sentence or its semantics can be described. Lemmatization will return the base or dictionary form of each lexical term or token based on its POS-tag. For example, the word “starting” which is a form of verb will return a “start” during lemmatization process. Words or lexical terms that share a common meaning will be grouped together in SentiWordNet lexicon to form a set, also known as synsets (synonym sets). Each lexical term will then be parsed to obtain its positive and negative scores characterized by the synsets. If the tokenized term is specified in the SentiWordNet lexicon, the polarity of the word can then be calculated by subtracting the negative score from the positive score according to the following:

$$\text{synset}_{\text{score}} = \text{positive}_{\text{score}} - \text{negative}_{\text{score}} \quad (1)$$

The overall score of the tweet can be calculated as:

$$\text{overall}_{\text{score}} = \frac{\sum \text{synset}_{\text{score}}}{\sum \text{tokens}} \quad (2)$$

The polarity categorization of the corresponding tweets can be obtained by comparing the sentiment score as shown below:

$$\text{overall}_{\text{score}} > 0 : \text{“positive”} \quad (3)$$

$$\text{overall}_{\text{score}} < 0 : \text{“negative”} \quad (4)$$

$$\text{overall}_{\text{score}} \equiv 0 : \text{“neutral”} \quad (5)$$

#### 3.3.2. TextBlob

The TextBlob library will return both polarity and subjectivity scores of each tweet. Polarity scores, which range from  $-1$  to  $+1$ , reflect the intensity of sentiment displayed by the authors through their tweets. Meanwhile, subjectivity scores, which range from  $0$  to  $1$ , indicate the views of the authors on the subject of discussion in which the value of  $0$  means that the tweet is very objective or displaying the factual information regarding the subject, and the value of  $1$  means that the tweet is very subjective and embodies the opinions of

the authors strongly. The polarity score obtained will be used to classify or categorize the tweets into sentiment classes according to the threshold below:

$$\text{polarity}_{\text{score}} > 0 : \text{"positive"} \quad (6)$$

$$\text{polarity}_{\text{score}} < 0 : \text{"negative"} \quad (7)$$

$$\text{polarity}_{\text{score}} \equiv 0 : \text{"neutral"} \quad (8)$$

### 3.3.3. VADER

VADER lexicon takes into account several characteristics of the input sentence during sentiment classification and will return the metric values containing negative, neutral, positive, and compound scores, or the "normalized weighted composite score". Every word within the text would be analyzed first to obtain the list of words that are present in VADER lexicon, and it will return their individual valence score. The valence score assigned to the words ranges from  $-4$  to  $+4$ , corresponding to the most negative and the most positive, respectively. The sum of all lexicon ratings has been normalized to be between the value of  $-1$  and  $+1$  and will be returned as a compound score which would be used as the threshold value for sentiment class categorization. The typical threshold value used is  $0.05$  but this study has opted for  $0.3$  as it gives the highest accuracy in the sentiment classification:

$$\text{Positive sentiment} : \text{compound}_{\text{score}} \geq 0.3 \quad (9)$$

$$\text{Negative sentiment} : \text{compound}_{\text{score}} \leq -0.3 \quad (10)$$

$$\text{Neutral sentiment} : \text{compound}_{\text{score}} < -0.3 \ \& \ \text{compound}_{\text{score}} > 0.3 \quad (11)$$

### 3.3.4. SentiStrength

SentiStrength lexicon can be accessed by using the SentiStrength2.3 tool downloaded for Windows. The program accepts a text file containing the list of texts as input and will report the positive score, negative score and emotion rationale that provides the sentiment score for each word in the analyzed text. SentiStrength detects the strength of the sentiment expressed and returns the value ranging from  $+1$  to  $+5$ , indicating 'not positive' to 'extremely positive', and  $-1$  to  $-5$  indicating 'not negative' to 'extremely negative'. The negative and positive scores can then be used for sentiment categorization as follows:

$$\text{Positive sentiment} : \text{positive}_{\text{score}} > \text{negative}_{\text{score}} \quad (12)$$

$$\text{Negative sentiment} : \text{positive}_{\text{score}} < \text{negative}_{\text{score}} \quad (13)$$

$$\text{Neutral sentiment} : \text{positive}_{\text{score}} \equiv \text{negative}_{\text{score}} \quad (14)$$

### 3.3.5. Hu and Liu Opinion Lexicon

Hu and Liu lexicon is composed of 6790 words that are divided into two dictionaries, each containing a list of words with positive polarity (2006 words) and negative polarity (4783 words). One of the characteristics of this lexicon is that it includes spelling mistakes that are common on social media platforms. Hu and Liu lexicon can be accessed by using the Orange tool and it will report the sentiment score of each analyzed piece of text, which would then be used as the threshold values for sentiment classification as follows:

$$\text{sentiment}_{\text{score}} > 0 : \text{"positive"} \quad (15)$$

$$\text{sentiment}_{\text{score}} < 0 : \text{"negative"} \quad (16)$$

$$\text{sentiment}_{\text{score}} \equiv 0 : \text{"neutral"} \quad (17)$$

### 3.3.6. MPQA Subjectivity Lexicon

MPQA lexicon has 2718 positive words and 4909 negative words in its dictionary. The MPQA dictionary includes words that have been tagged individually as adjectives, adverbs, anypos (any part-of-speech), nouns and verbs that can also be either weakly or strongly subjective. In this study, the sentiment analysis of climate change tweets by using MPQA lexicon was conducted using the Orange tool. Only words that have negative or positive sentiment polarity in the MPQA dictionary will be used during the sentiment classification task. During the process, every word in the tweets that exists within the dictionary will be tagged with either positive or negative sentiment and the average sentiment score will be calculated by counting the total sentiment polarity divided by the total number of sentiment words. The positive and negative dictionary for MPQA can be downloaded from Wilson et al. [55]. The sentiment score will be used as threshold values for sentiment classification, as in Equations (15)–(17) above.

### 3.3.7. WKWSCI Lexicon

WKWSCI is a general-purpose sentiment lexicon that was manually annotated by undergraduate students from Nanyang Technological University, Singapore. It was constructed based on 12 source dictionaries containing common American English word lists and every word has been coded into either positive or negative sentiment polarity [18]. Each word with sentiment polarity was then coded into several subcategories reflecting its sentiment polarity strength, which are slightly positive (sentiment value of 1), positive (2), very positive (3), slightly negative (−1), negative (−2) and very negative (−3). WKWSCI comprises of 3121 positive words and 7100 negative words that have been POS-tagged with adjectives, adverbs, verbs, and nouns. The adaptation of WKWSCI into this study follows the same steps as the implementation of MPQA lexicon whereby only words with positive and negative tags have been used during the sentiment classification task. The positive and negative dictionary for WKWSCI can be downloaded from Khoo and Johnkhan [10]. The sentiment score obtained will be used for sentiment polarity categorization of each tweet, as in Equations (15)–(17) above.

## 3.4. Feature Extraction Technique

Two feature extraction techniques have been adopted in this study which are Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF–IDF). These two feature extraction techniques have been chosen due to some studies indicate that LR, SVM and NB achieved higher accuracies using BoW compared to TF–IDF. The feature extraction process is important for sentiment classification using machine learning approaches as large feature spaces can increase the complexity and reduce the ability to solve the tasks [56].

### 3.4.1. Bag-of-Words (BoW)

A feature vector will be built next based on this vocabulary to be used during the training of the machine learning algorithms. The feature vector takes the form of  $d = (x_1, x_2, \dots, x_l)$ , where  $x_i$  denotes the number of occurrences of the word term in the text. It works on the basis of the presence or absence of the word term in the whole dictionary or document [57]. For example, the whole dictionary consists of the terms {"global", "warming", "is", "bad", "for", "environment"}, meanwhile the first sentence is  $d_1$  "global warming and climate change" and the second sentence is  $d_2$  "excess CO<sub>2</sub> in environment is bad". In BoW representation,  $d_1$  and  $d_2$  will be projected to be a vector of  $d_1 = (1, 1, 0, 0, 0, 0)$  and  $d_2 = (0, 0, 0, 1, 1, 1)$  which indicate the limitations of this method to be the sparsity of the vectors produced and the inability to capture the semantic orientation of the text [57].

### 3.4.2. Term Frequency–Inverse Document Frequency (TF–IDF)

This technique involves a comparison between the words in the documents and their relevancy within the overall documents. The term frequency (TF) as the name suggests,

measures the number of times the word or term appears in a document and depends on the length of that document. The equation to calculate TF value is given in Equation (18). For example, if the document contains 20 words with the term “global” appearing 3 times, the TF value for the word “global” would be  $TF = \frac{3}{20} = 0.15$ :

$$TF(i, j) = \frac{\text{Term } i \text{ frequency in document } j}{\text{Total words in document } j} \quad (18)$$

Meanwhile the inverse document frequency (IDF) measures how common the words are in the overall documents which will give us the extent of how informative they are or the weights for model training [46]. The equation to calculate IDF value is given in Equation (19). For example, if the total number of documents available is 250 and 150 of them contain the term “global”, the IDF value for word “global” would therefore be  $IDF = \log \frac{250}{150} = 0.22$ :

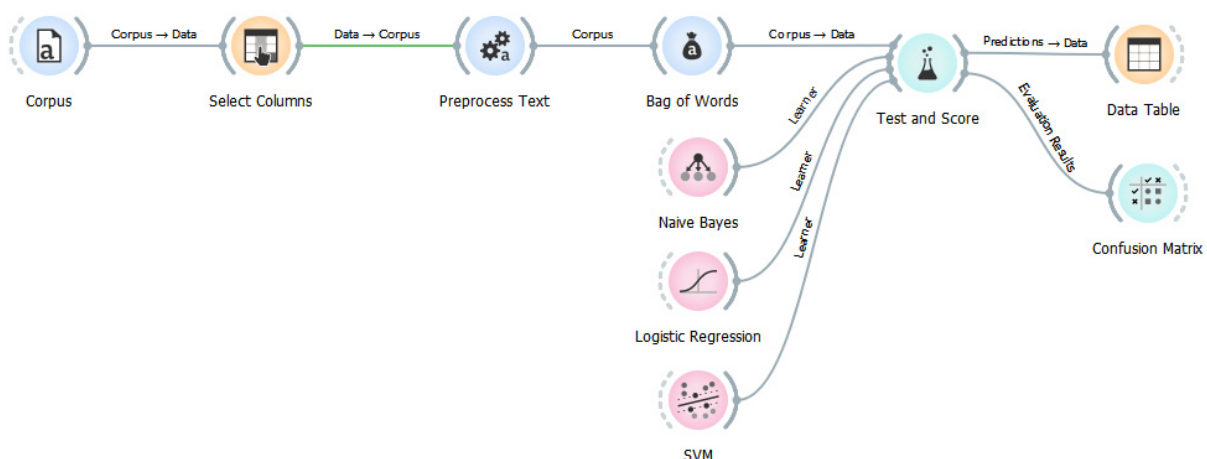
$$IDF(i) = \log \frac{\text{Total documents}}{\text{documents with term } i} \quad (19)$$

TF-IDF is the product of multiplying these two frequency metrics together and its objective is to give low weights to the terms that often appear in the overall documents [58]. The equation to calculate TF-IDF value is given in Equation (20). Therefore, the TF-IDF value for the term “global” would be  $TF - IDF = 0.15 \times 0.22 = 0.033$ :

$$TF - IDF = TF \times IDF \quad (20)$$

### 3.5. Supervised Machine Learning Methods

Three supervised machine learning methods have been adopted in this study, which are Logistic Regression (LR), Support Vector Machine (SVM) and Naïve Bayes (NB). These three machine learning classifiers have been chosen due to them being in the top 3 most utilized machine learning classifiers in the sentiment analysis field. The training of machine learning algorithms will employ the cross-validation method with 10-folds, as it can reduce the variance to obtain the performance of the models and will be carried out by using the Orange data mining tool, as shown in Figure 2. The input to the machine learning algorithms would be the feature vectors defined earlier, which are BoW and TF-IDF; meanwhile, the target variable is the original polarity score of the tweets that have been determined by the annotators.



**Figure 2.** Sentiment analysis workflow for the implementation of machine learning classifiers by using the Orange data mining tool.

#### 3.5.1. Logistic Regression

Logistic regression, which is a type of exponential or log-linear classifier, tries to find the relationship between a categorical dependent variable, which is the sentiment polarity

of the tweets, and one or more independent variables, which are the features that have been extracted using BoW and TF-IDF. A dictionary containing a list of all the words available in the dataset will be constructed together with their frequencies of occurrences in the three sentiment classes, which are negative, neutral, and positive. The final probability is calculated by using the softmax function, where it will turn the probability to range from 0 to 1 and the sum of all the probabilities will be equal to 1. The output of the logistic regression is the calibrated probability of the tweet belonging to each sentiment class.

### 3.5.2. Support Vector Machine

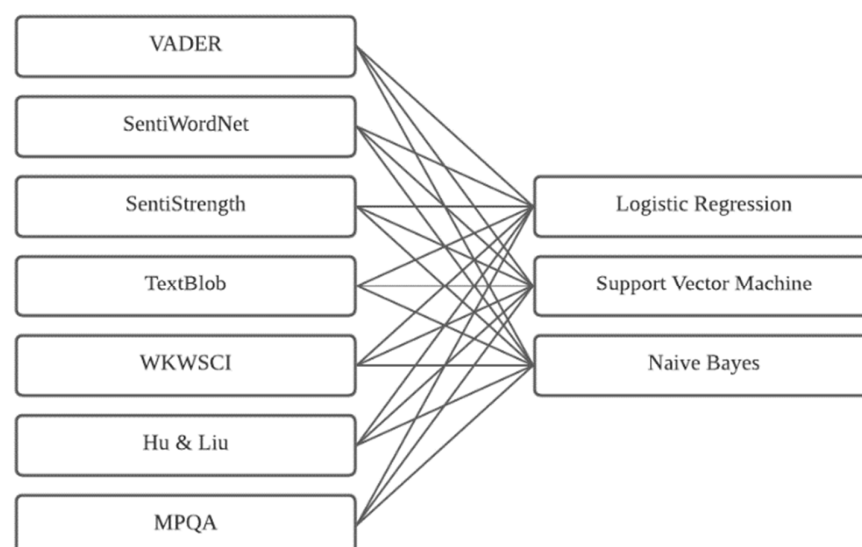
Support vector machine algorithms try to find the hyperplane that maximizes the distance of the support vectors or the margin between two classes. This hyperplane or line boundary will separate the data into groups, which in this study are positive, neutral, and negative. The kernel function, which projects the support vectors from low-dimensional space to higher-dimensional space, is an important parameter for training and this study uses linear kernel because it yields a better accuracy compared to other kernel functions, as has been shown by Ramasamy et al. [59].

### 3.5.3. Naïve Bayes

Naïve Bayes algorithm, which is a conditional probability model, applies the Bayes theorem with strong (Naïve) independence assumptions which means that every feature is independent of other features in order to determine the probability of those features in the corpus [7].

## 3.6. Hybrid Methods

For the hybridization between unsupervised lexicon-based and supervised machine learning-based approaches, the polarity score of every tweet given by each lexicon will be used as a target variable during the training of machine learning classifiers. Figure 3 shows the combination of lexicon and machine learning classifiers in hybrid methods. Each lexicon will be hybridized between each machine learning classifier. This training of hybrid models will be carried out by using the Orange data mining tool for all the classifiers, as shown in Figure 2. The output of the machine learning classifiers will be the probability of the text belonging in each sentiment polarity class. The highest probability will be selected as representing the polarity of the text.



**Figure 3.** The combination of lexicon and machine learning classifiers in hybrid models.

The use of artificial intelligence in many areas has shown considerable improvement in the performance and accuracy in its precision [60], prediction [61] and computer assisted systems [62,63].

### 3.7. Measurement Metrics

In evaluating the performance of each approach, several evaluation parameters will be used, namely accuracy, precision, recall, and F1-score:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (21)$$

Accuracy measures the fraction of correct predictions and has the limitation of being an incorrect or misleading measurement when imbalanced datasets are used. TP denotes the true positive, TN is the true negative, FP is the false positive and FN is the false negative values. Other evaluation parameters, which are precision, recall and F1-score, will also be used to 1 the shortcomings of accuracy measurement:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (22)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (23)$$

Precision reports the performance of the classifiers in correctly predicting all the positive classes for all positive predictions. It is the fraction of true positive in all predicted positive classes. Meanwhile, recall reports the performance of the classifiers in correctly predicting all positive classes for all actual positive values. It is the fraction of true positive in all actual positive classes. F1-score is the harmonic mean of precision and recall which quantifies the performance evaluation parameters into one value or score:

$$\text{F1}_{\text{score}} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \quad (24)$$

## 4. Results and Discussion

The primary components of this section will cover the results obtained through the implementation of the classification models. The first section is the construction of lexicon-based models. The second section is the training and evaluation of the machine learning-based approach by using the original annotated polarities. The third section is the construction of the hybrid approach. A summary on the discussions of the results obtained is provided at the end of this section which highlights the main findings of this study.

### 4.1. Lexicon-Based Approaches

The results for the comparison between seven different lexicons will be presented in this section. For Earth Hour 2015 Corpus, the highest performing lexicon is Hu and Liu with an accuracy of 96.0%; meanwhile, the lowest performing lexicon is SentiWordNet with an accuracy of 50.0% for non-lemmatized texts, as shown in Table 2. This depicts that there is a huge discrepancy in the performances between the lexicons. The same results have been obtained for lemmatized texts, with Hu and Liu and SentiWordNet achieving 90.5% and 49.6%, respectively, as shown in Table 3. Meanwhile, for Weather Sentiment the highest performing lexicon is SentiStrength with an accuracy of 59.6%, and the lowest performing lexicon is SentiWordNet with an accuracy of 47.8% for non-lemmatized texts. The same results have been obtained for lemmatized texts with SentiStrength and SentiWordNet achieving 58.8% and 48.0%, respectively.

**Table 2.** The performance of the lexicons without lemmatization for all datasets.

Dataset	Lexicon	Accuracy	Precision	Recall	F1-Score
Earth Hour 2015 Corpus	VADER	73.0	62.6	59.8	59.5
	SentiWordNet	50.0	46.0	52.3	45.6
	TextBlob	67.1	58.9	60.4	59.6
	SentiStrength	70.2	56.8	57.7	57.0
	Hu and Liu	96.0	94.6	95.9	95.3
	MPQA	73.0	66.8	74.9	69.1
	WkWSCI	60.3	55.4	62.7	55.8
Weather Sentiment	VADER	59.0	60.3	60.3	58.0
	SentiWordNet	47.8	48.0	48.6	46.3
	TextBlob	57.7	60.4	58.6	57.0
	SentiStrength	59.6	59.7	60.5	59.5
	Hu and Liu	57.4	58.4	58.2	57.3
	MPQA	48.8	48.9	49.8	48.1
	WkWSCI	53.6	53.6	54.6	53.3
Climate Change Sentiment	VADER	47.2	48.1	45.8	50.0
	SentiWordNet	36.1	42.1	38.6	32.3
	SentiStrength	39.9	46.7	46.7	39.3
	Hu and Liu	49.2	53.5	54.2	49.1
	MPQA	52.0	51.3	52.7	50.8
	WkWSCI	50.3	52.0	52.5	49.6
	Combined Dataset	VADER	57.2	58.8	57.2
SentiWordNet		44.9	47.0	46.5	44.3
TextBlob		71.2	72.6	69.7	70.0
SentiStrength		56.0	55.8	55.2	55.3
Hu and Liu		62.0	61.7	60.9	61.1
MPQA		54.0	52.9	53.0	52.8
WkWSCI		53.9	53.3	53.4	53.3

**Table 3.** The performance of the lexicons with lemmatization for all datasets.

Dataset	Lexicon	Accuracy	Precision	Recall	F1-Score
Earth Hour 2015 Corpus	VADER	71.8	60.2	59.1	58.5
	SentiWordNet	49.6	46.9	55.5	46.3
	TextBlob	54.0	52.5	55.2	52.5
	SentiStrength	70.2	56.8	57.7	57.0
	Hu and Liu	90.5	87.4	90.9	88.8
	MPQA	60.3	61.3	67.8	59.8
	WkWSCI	67.5	62.1	68.6	63.4
Weather Sentiment	VADER	58.2	59.6	59.6	57.2
	SentiWordNet	48.0	48.0	48.7	46.6
	TextBlob	57.1	59.8	58.1	56.5
	SentiStrength	58.8	59.0	59.7	58.8
	Hu and Liu	57.1	57.5	57.9	57.2
	MPQA	50.7	50.6	51.8	50.2
	WkWSCI	54.4	54.0	55.4	54.1
Climate Change Sentiment	VADER	45.7	44.8	44.3	38.7
	SentiWordNet	35.6	38.1	37.6	31.4
	SentiStrength	38.6	44.8	44.3	38.1
	Hu and Liu	47.0	49.9	50.9	46.8
	MPQA	54.0	52.7	54.2	52.5
	WkWSCI	49.5	51.8	52.1	49.1
	Combined Dataset	VADER	55.8	57.4	55.7
SentiWordNet		44.8	46.9	46.3	44.2
TextBlob		65.6	67.9	64.6	64.7
SentiStrength		55.2	54.9	54.3	54.4
Hu and Liu		60.2	59.7	59.5	59.5
MPQA		53.4	52.6	52.9	52.3
WkWSCI		55.4	54.8	55.2	54.8



The Climate Change Sentiment dataset was originally annotated by using TextBlob; therefore, the performance obtained by using this lexicon has been left out from further analysis in order to avoid the bias posed. The highest performing lexicon is MPQA with an accuracy of 52.0% and the lowest performing lexicon is SentiWordNet with an accuracy of 36.1% for non-lemmatized texts. The same results have been obtained for lemmatized texts, with MPQA and SentiWordNet achieving 54.0% and 35.6%, respectively. The accuracy that has been achieved by each of the lexicons between lemmatized and non-lemmatized texts for both Weather Sentiment and Climate Change Sentiment did not divert too much from each other, unlike the Earth Hour 2015 Corpus. This concludes the stability of both datasets in investigating the performances of sentiment analysis approaches in classifying climate change tweets.

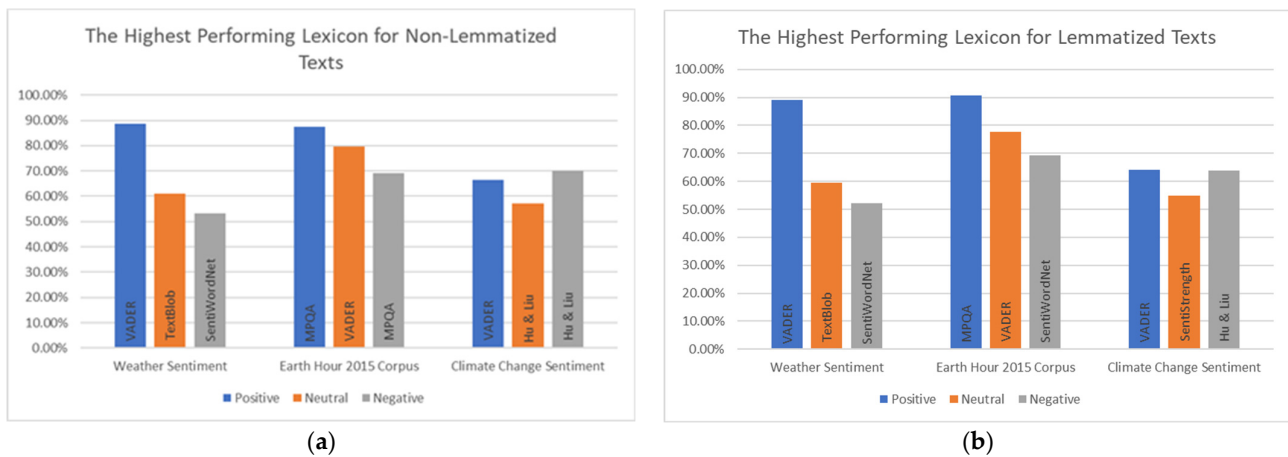
Since one of the datasets, which is Climate Change Sentiment, within the Combined Dataset has been fully annotated by TextBlob, this lexicon will be omitted from further analysis. Furthermore, since Hu and Liu also posed a potential bias towards the Earth Hour 2015 Corpus, this lexicon might as well be omitted from further analysis. These two lexicons can be seen to dominate the first and second place in the highest performing lexicon for the Combined Dataset. Therefore, the highest performing lexicon for Combined Dataset is VADER with an accuracy of 57.2% and the lowest performing lexicon is SentiWordNet with an accuracy of 44.9% for non-lemmatized texts. The same results have been obtained for lemmatized texts with VADER and SentiWordNet achieving 55.8% and 44.8%, respectively.

One striking observation is the low average accuracy of below 60% that has been attained by the highest performing lexicons on each dataset, except for Earth Hour 2015 Corpus, which could be due to the potential bias of this dataset on the Hu and Liu lexicon as shown in Table 4. Since the majority of the lexicons score higher on Recall compared to Precision overall on each dataset, it can be concluded that all lexicons overclassify neutral tweets as opinionated, i.e., either positive or negative regardless of whether the tweets are lemmatized or otherwise.

**Table 4.** The summary of the best performing lexicon for each dataset with different preprocessing techniques.

Dataset	Data Preprocessing	Lexicon	Accuracy
Earth Hour 2015 Corpus	With Lemmatization	Hu and Liu	90.5
	Without Lemmatization	Hu and Liu	96.0
Weather Sentiment	With Lemmatization	SentiStrength	58.8
	Without Lemmatization	SentiStrength	59.6
Climate Change Sentiment	With Lemmatization	MPQA	54.0
	Without Lemmatization	MPQA	52.0
Combined Dataset	With Lemmatization	VADER	55.8
	Without Lemmatization	VADER	57.2

For the other two datasets, which are Weather Sentiment and Climate Change Sentiment, the accuracy of all the lexicons hovers around 60%. However, for Earth Hour 2015 Corpus, there is a 40.9% difference between the highest and lowest performing lexicons which are Hu and Liu and SentiWordNet, respectively. This might reflect the large discrepancy between the tweets and the sentiment polarity given by the annotators for this specific dataset. For both non-lemmatized and lemmatized texts, VADER performed the best in classifying positive and neutral tweets, meanwhile SentiWordNet dominates the classification of negative tweets. This can be seen in Figure 4a,b.



**Figure 4.** (a) The highest performing lexicon for non-lemmatized texts; (b) the highest performing lexicon for lemmatized texts for each sentiment class.

The analysis shows that there was very little confusion between negative and positive for all the lexicons and not much confusion between negative misclassification from neutral as shown in Table 5, and this confusion has been addressed by Maynard and Bontcheva [50]. Since the distinctions between negative with positive and negative with neutral are the most important factors to be clear about in investigating the sentiment of climate change and environmental related tweets, it can be concluded that all the lexicons seem to have achieved this objective, except for VADER, SentiWordNet and SentiStrength. VADER and SentiWordNet have the tendency to flip the polarity from positive to negative and vice versa, meanwhile SentiStrength has the tendency to misclassify negative tweets into neutral. Therefore, the results show that most of the lexicons are appropriate to be adapted in this task of classifying climate change related tweets.

**Table 5.** The bias of the lexicons towards other sentiment classes on non-lemmatized texts for all datasets. For example, “Neg to Pos” means that the lexicon tends to misclassify negative tweets into the positive sentiment class.

Dataset	Bias	VADER	SentiWordNet	Senti Strength	TextBlob	Hu and Liu	MPQA	WKW SCI
Earth Hour 2015 Corpus	Neg to Pos Pos to Neg Neg to Neu			✓				
Weather Sentiment	Neg to Pos Pos to Neg Neg to Neu	✓	✓	✓		✓	✓	✓
Climate Change Sentiment	Neg to Pos Pos to Neg Neg to Neu	✓	✓					
Combined Dataset	Neg to Pos Pos to Neg Neg to Neu	✓	✓	✓		✓	✓	✓

#### 4.2. Machine Learning-Based Approaches

In general, the performance of Logistic Regression (LR) is higher than both Support Vector Machine (SVM) and Naïve Bayes (NB). Meanwhile, there is no evidence that SVM is better than NB as they depend on which dataset the machine learning classifier was being trained on. It can also be noted that the performance of LR did not divert too much from each dataset, unlike those obtained by SVM and NB.

Furthermore, the performance of machine learning classifiers that were trained using the BoW feature extraction technique is higher than TF-IDF in some of the datasets, such as Weather Sentiment and Combined Dataset, as shown in Table 6. However, these feature extraction techniques did not have any effect of increasing nor decreasing the performance of SVM and NB for both lemmatized and non-lemmatized texts, as shown in Table 7. Moreover, it can also be seen that increasing the number of training examples in the classifiers did not increase the performance of the models as anticipated. This could be attributed to the discrepancy in the annotation process because each dataset was annotated by different annotators. Lemmatization also improves the performance of the machine learning classifiers compared to the non-lemmatized texts.

**Table 6.** The performance of the machine learning approach trained using lemmatized texts.

Dataset	Feature Extraction	Logistic Regression				Support Vector Machine				Naïve Bayes			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Earth Hour 2015 Corpus	BoW	72.6	72.0	72.6	69.4	76.6	75.7	76.6	74.9	48.4	69.8	48.4	53.8
	TF-IDF	74.6	76.2	74.6	71.5	76.6	75.7	76.6	74.9	48.4	69.8	48.4	53.8
Weather Sentiment	BoW	74.7	74.7	74.7	74.7	70.4	70.5	70.4	70.4	72.2	73.6	72.2	72.4
	TF-IDF	74.3	74.4	74.3	74.3	70.4	70.5	70.4	70.4	72.2	73.6	72.2	72.4
Climate Change Sentiment	BoW	63.4	62.2	63.4	62.3	60.9	61.0	60.9	60.4	46.2	61.3	46.2	42.0
	TF-IDF	64.4	65.2	64.4	62.6	60.9	61.0	60.9	60.4	46.2	61.3	46.2	42.0
Combined Dataset	BoW	70.2	70.2	70.2	70.0	55.8	57.1	55.8	55.9	63.1	63.9	63.1	62.6
	TF-IDF	68.7	68.9	68.7	68.4	55.8	57.1	55.8	55.9	63.1	63.9	63.1	62.6

**Table 7.** The performance of the machine learning approach trained using non-lemmatized texts.

Dataset	Feature Extraction	Logistic Regression				Support Vector Machine				Naïve Bayes			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Earth Hour 2015 Corpus	BoW	73.4	73.1	73.4	70.4	75.4	75.4	75.4	73.4	46.0	68.2	46.0	51.4
	TF-IDF	74.6	75.7	74.6	71.0	74.6	74.5	74.6	72.5	46.0	68.2	46.0	51.4
Weather Sentiment	BoW	72.9	73.0	72.9	72.9	70.0	70.0	70.0	70.0	72.5	73.7	72.5	72.6
	TF-IDF	72.9	72.9	72.9	72.8	70.0	70.0	70.0	70.0	72.5	73.7	72.5	72.6
Climate Change Sentiment	BoW	62.6	61.5	62.6	61.5	61.9	62.4	61.9	61.2	46.7	65.2	46.7	42.2
	TF-IDF	61.9	62.6	61.9	59.9	61.9	62.4	61.9	61.2	46.7	65.2	46.7	42.2
Combined Dataset	BoW	69.2	69.3	69.2	69.1	54.1	55.5	54.1	54.2	62.6	64.0	62.6	62.0
	TF-IDF	68.1	68.3	68.1	67.8	54.1	55.5	54.1	54.2	62.6	64.0	62.6	62.0

### 4.3. Hybrid Approaches

The results for the hybrid approach will be presented in this section. In the hybrid approach, the sentiment polarities given by the lexicons were used as the target variable during the training of the machine learning classifiers. The F1-score will be given a higher weight against the other measurement metrics in comparing the performance of the classification models due to the imbalanced nature of the datasets.

#### 4.3.1. Hybrid Approach for Lemmatized Texts

As for the results of BoW in Earth Hour 2015 Corpus, TextBlob had the highest performance when it had been hybridized with LR (74.2%), SVM (74.2%), and NB (64.8%) as shown in Table 8. As for the results of TF-IDF, Hu and Liu had the highest performance with LR (72.5%) while TextBlob had the highest performance with SVM (74.2%) and NB (64.8%).

As for the results of BoW in Weather Sentiment, TextBlob had the highest performance when it had been hybridized with all machine learning classifiers, namely LR (72.9%), SVM (70.5%), and NB (64.7%). As for the results of TF-IDF, the same outcome had been achieved as with BoW, i.e., TextBlob had the highest performance with LR (73.1), SVM (70.5%) and NB (64.7%).

**Table 8.** The performance (F1-score) of the hybrid approach for lemmatized texts.

Dataset	Feature Extraction	BoW			TF-IDF		
	Classifiers	LR	SVM	NB	LR	SVM	NB
Earth Hour 2015 Corpus	VADER	71.5	73.9	49.5	71.7	74.2	49.5
	SentiWordNet	62.5	62.4	57.9	64.3	62.4	57.9
	TextBlob	74.2	74.2	64.8	71.2	74.2	64.8
	SentiStrength	71.9	72.8	57.2	70.6	72.8	57.2
	Hu and Liu	71.2	74.2	59.4	72.5	72.9	59.4
	MPQA	66.6	69.3	62.6	68.8	70.1	62.6
	WKWSCI	63.9	62.9	52.8	63.1	62.5	52.8
Weather Sentiment	VADER	69.0	65.3	63.2	69.3	65.3	63.2
	SentiWordNet	65.9	57.3	43.7	64.4	57.3	43.7
	TextBlob	72.9	70.5	64.7	73.1	70.5	64.7
	SentiStrength	60.7	57.6	60.0	60.6	57.6	60.0
	Hu and Liu	63.8	59.6	60.2	63.1	59.6	60.2
	MPQA	62.0	56.0	60.7	62.3	56.0	60.7
	WKWSCI	63.0	59.5	61.4	64.3	59.5	61.4
Climate Change Sentiment	VADER	63.2	66.0	5.9	64.3	66.0	5.9
	SentiWordNet	62.4	61.0	6.6	63.6	61.0	6.6
	TextBlob	66.9	62.9	33.0	65.2	62.9	33.0
	SentiStrength	53.0	54.4	32.5	53.6	54.4	32.5
	Hu and Liu	51.9	52.8	46.2	51.0	52.8	46.2
	MPQA	56.8	54.2	42.3	53.9	54.2	42.3
	WKWSCI	55.0	50.9	42.0	53.4	50.9	42.0
Combined Dataset	VADER	71.8	61.0	51.5	72.4	61.0	51.5
	SentiWordNet	66.8	51.9	44.3	63.4	51.9	44.3
	TextBlob	74.7	61.8	51.9	75.3	61.8	51.9
	SentiStrength	66.0	54.7	59.5	65.7	54.7	59.5
	Hu and Liu	67.1	52.9	60.2	66.9	52.9	60.2
	MPQA	65.6	53.2	55.8	64.9	53.2	55.8
	WKWSCI	65.3	51.9	58.7	66.4	51.9	58.7

As for the results of BoW in Climate Change Sentiment, the combinations that had the highest performance were TextBlob with LR (66.9%), VADER with SVM (66.0%), and Hu and Liu with NB (46.2%). As for the results of TF-IDF, TextBlob had the highest performance when it had been hybridized with LR (65.2%) and VADER with SVM (66.0%). Meanwhile, for the NB classifier, the highest performance was achieved by Hu and Liu (46.2%).

As for the results of BoW in the Combined Dataset, TextBlob had the highest performance when it had been hybridized with LR (74.7%) and SVM (61.8%), while for NB, the combination between this classifier with Hu and Liu yielded the highest performance with an accuracy of 60.2%. As for the results of TF-IDF, the same outcome had been achieved as with BoW, i.e., TextBlob had the highest performance with LR (75.3%) and SVM (61.8%), while Hu and Liu yielded the highest accuracy for NB with 60.2%.

#### 4.3.2. Hybrid Approach for Non-Lemmatized Texts

As for the results of BoW in the Earth Hour 2015 Corpus, TextBlob had the highest performance for predicting Earth Hour related tweets when it had been hybridized with LR (70.0%) and SVM (75.3%) with WKWSCI had the highest performance with NB (60.6%) as shown in Table 9. As for the results of TF-IDF, the combination that had the highest performance are TextBlob with LR (70.4%) and SVM (74.8%), while for NB, WKWSCI had the highest performance (60.6%).

**Table 9.** The performance (F1-score) of the hybrid approach for non-lemmatized texts.

Dataset	Feature Extraction	BoW			TF-IDF		
	Classifiers	LR	SVM	NB	LR	SVM	NB
Earth Hour 2015 Corpus	VADER	67.2	70.6	43.7	69.8	71.1	43.7
	SentiWordNet	56.0	57.9	52.2	57.7	57.2	52.2
	TextBlob	70.0	75.3	58.2	70.4	74.8	58.2
	SentiStrength	66.2	69.8	52.2	69.7	70.2	52.2
	Hu and Liu	67.1	74.5	56.2	70.0	74.0	56.2
	MPQA	62.9	67.4	58.4	67.7	67.4	58.4
	WKWSCI	65.5	61.5	60.6	63.4	61.5	60.6
Weather Sentiment	VADER	69.2	64.2	61.3	69.1	64.2	61.3
	SentiWordNet	65.6	58.3	44.5	65.1	58.3	44.5
	TextBlob	74.8	67.2	62.8	74.1	67.2	62.8
	SentiStrength	62.6	55.4	58.4	60.6	55.4	58.4
	Hu and Liu	61.9	59.9	59.7	61.8	59.9	59.7
	MPQA	61.8	55.9	57.8	61.8	55.9	57.8
	WKWSCI	61.7	57.7	57.5	63.5	57.7	57.5
Climate Change Sentiment	VADER	62.9	66.2	0.5	64.0	66.2	0.5
	SentiWordNet	65.9	63.7	3.2	65.4	63.7	3.2
	TextBlob	61.5	61.2	42.2	59.9	61.2	42.2
	SentiStrength	56.5	56.4	28.2	56.3	56.4	28.2
	Hu and Liu	52.3	51.4	45.0	51.0	51.4	45.0
	MPQA	54.2	52.7	44.5	52.9	52.7	44.5
	WKWSCI	51.5	50.7	50.1	50.4	50.4	50.1
Combined Dataset	VADER	71.8	61.3	49.7	71.8	61.3	49.7
	SentiWordNet	67.0	53.4	43.4	65.8	53.4	43.4
	TextBlob	74.3	59.9	57.8	73.7	59.8	57.8
	SentiStrength	64.6	52.1	58.1	63.0	52.1	58.1
	Hu and Liu	65.4	52.8	59.2	65.4	52.9	59.2
	MPQA	64.6	49.5	55.4	63.2	49.5	55.4
	WKWSCI	63.0	51.6	59.6	62.6	50.6	59.6

As for the results of BoW in Weather Sentiment, TextBlob had the highest performance for predicting weather related tweets when it had been hybridized with all three machine learning classifiers, which were LR (74.8%), SVM (67.2%) and NB (62.8%). As for the results of TF-IDF, the same conclusion as the BoW had been achieved, whereby TextBlob combined with all three machine learning classifiers had the highest performance which were LR (74.1%), SVM (67.2%) and NB (62.8%).

As for the results of BoW in Climate Change Sentiment, the combinations that had the highest performance were SentiWordNet with LR (65.9%), VADER with SVM (66.2%) and WKWSCI with NB (50.1%). As for the results of TF-IDF, the combinations that had the highest performance for classifying climate change related discussions were SentiWordNet with LR (65.4%), VADER with SVM (66.2%), and WKWSCI with NB (50.1%), which is the same as had been obtained through the BoW feature extraction technique.

As for the results of BoW in the Combined Dataset, the combinations that had the highest performance were TextBlob with LR (74.3%), VADER with SVM (61.3%), and WKWSCI with NB (59.6%). As for the results of TF-IDF, TextBlob had the highest performance when it had been hybridized with LR (73.7%), while for SVM it was VADER (61.3%). Meanwhile, for the NB classifier, WKWSCI achieved the highest performance with an F1-score of 59.6%.

For the Earth Hour 2015 Corpus, it can be noted that the performance of each combination did not divert too much from each other, for example, the F1-score hovers around 60% to 70% for LR and SVM. Furthermore, the difference in the performance between the highest and the lowest for LR is only 14%, unlike that obtained previously by lexicon-based

approaches, which is around 50%. This shows that hybrid approaches compensate for the discrepancy in the original annotation of this dataset, i.e., between the annotators.

The reason behind combining all the datasets together is to investigate the effect of increasing the number of training examples on the performances obtained by the machine learning classifiers. Therefore, the performances obtained by the Combined Dataset have been compared against those yielded by individual datasets. However, this study witnessed no significant increase in the performances of the machine learning classifiers when the number of training examples is bigger. This finding confirms the study by Abdelwahab et al. [64], which stated that changing the training size has no significant effect on the performance of sentiment classification by using machine learning approaches.

It can be seen that for the majority of the machine learning classifiers, TextBlob dominated the highest performance (Table 10). The results show that Logistic Regression outperformed both Support Vector Machine and Naïve Bayes. The results also indicate that the model using lemmatized texts as the training features was better than the model using non-lemmatized texts. Through the analysis of each dataset on the hybrid approaches, there are several lexicons that decreased the performance of the models when they were trained by using the lemmatized texts as compared to non-lemmatized texts. However, there is an insignificant number of lexicons that are exhibiting this behavior. Therefore, it can be concluded that lemmatization improved the performance of hybrid approaches.

**Table 10.** The summary of the highest performing hybrid approaches for each dataset that had been trained by using BoW and TF-IDF.

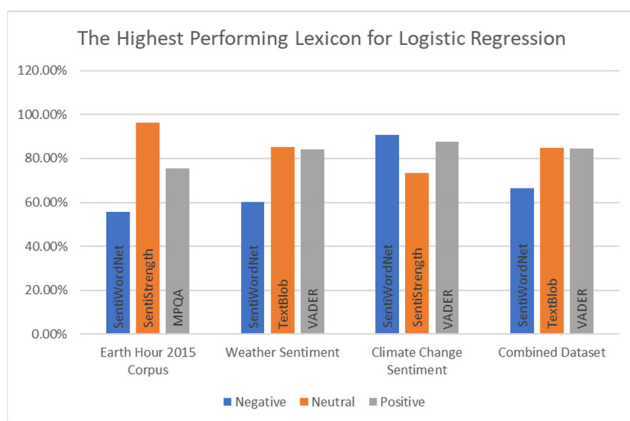
Dataset	Feature Extraction	BoW			TF-IDF		
		Classifiers	LR	SVM	NB	LR	SVM
Earth Hour 2015 Corpus	With Lemmatization	TextBlob (74.2%)	TextBlob (74.2%)	TextBlob (64.8%)	Hu and Liu (72.5%)	TextBlob (74.2%)	TextBlob (64.8%)
	Without Lemmatization	TextBlob (70.0%)	TextBlob (75.3%)	WkWSCI (60.6%)	TextBlob (70.4%)	TextBlob (74.8%)	WkWSCI (60.6%)
Weather Sentiment	With Lemmatization	TextBlob (72.9%)	TextBlob (70.5%)	TextBlob (64.7%)	TextBlob (73.1%)	TextBlob (70.5%)	TextBlob (64.7%)
	Without Lemmatization	TextBlob (74.8%)	TextBlob (67.2%)	TextBlob (62.8%)	TextBlob (74.1%)	TextBlob (67.2%)	TextBlob (62.8%)
Climate Change Sentiment	With Lemmatization	TextBlob (66.9%)	VADER (66.0%)	Hu and Liu (46.2%)	TextBlob (65.2%)	VADER (66.0%)	Hu and Liu (46.2%)
	Without Lemmatization	SentiWordNet (65.9%)	VADER (66.2%)	WkWSCI (50.1%)	SentiWordNet (65.4%)	VADER (66.2%)	WkWSCI (50.1%)
Combined Dataset	With Lemmatization	TextBlob (74.7%)	TextBlob (61.8%)	Hu and Liu (60.2%)	TextBlob (75.3%)	TextBlob (61.8%)	Hu and Liu (60.2%)
	Without Lemmatization	TextBlob (74.3%)	VADER (61.3%)	WkWSCI (59.6%)	TextBlob (73.7%)	VADER (61.3%)	WkWSCI (59.6%)

As discussed earlier, the distinctions between negative with positive and negative with neutral are the most important factors to be clear about in investigating the sentiment of climate change and environmental related tweets. By taking into account the first factor, VADER and SentiWordNet show signs of bias by favoring positive misclassification more than neutral from negative sentiment polarity (Table 11). This holds true for both Logistic Regression and Support Vector Machine. For the second factor, SentiStrength, Hu and Liu and WkWSCI that were combined with Logistic Regression showed bias towards neutral misclassification from negative. Hence, by considering the first and second factors, VADER, SentiWordNet, SentiStrength, Hu and Liu and WkWSCI that have been combined with Logistic Regression, are not suitable to be used in this sentiment analysis task.

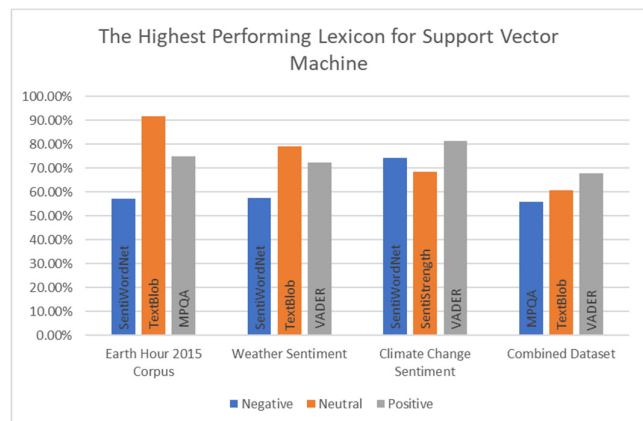
**Table 11.** The bias of the lexicons towards neutral and positive misclassification from negative using lemmatized texts and TF-IDF for all machine learning classifiers.

Machine Learning Classifier	Dataset	Earth Hour 2015 Corpus		Weather Sentiment		Climate Change Sentiment		Combined Dataset		
		Bias	Neg to Neu	Neg to Pos	Neg to Neu	Neg to Pos	Neg to Neu	Neg to Pos	Neg to Neu	Neg to Pos
Logistic Regression	VADER		✓			✓				✓
	SentiWordNet		✓			✓				✓
	TextBlob		✓				✓			
	SentiStrength		✓		✓		✓		✓	
	Hu and Liu		✓		✓		✓		✓	
	MPQA				✓				✓	
	WKWSCl		✓		✓		✓		✓	
Support Vector Machine	VADER		✓				✓			✓
	SentiWordNet		✓				✓			
	TextBlob									
	SentiStrength		✓		✓		✓		✓	
	Hu and Liu				✓		✓		✓	
	MPQA				✓		✓			
Naïve Bayes	VADER				✓		✓			✓
	SentiWordNet		✓		✓		✓		✓	
	TextBlob						✓		✓	
	SentiStrength				✓			✓		
	Hu and Liu				✓				✓	
	MPQA				✓		✓		✓	
	WKWSCl				✓		✓		✓	

It can be concluded that for both Logistic Regression and Support Vector Machine, SentiWordNet, TextBlob and VADER achieved the highest correct classification rate for classifying negative, neutral and positive tweets respectively as shown in Figure 5a,b.



(a)



(b)

**Figure 5.** (a) The highest performing lexicon for Logistic Regression with TF-IDF and lemmatized texts; (b) the highest performing lexicon for Support Vector Machine with TF-IDF and lemmatized texts.

#### 4.4. Discussion on All of the Approaches

The results indicate that all models are sensitive to the effect of lemmatization which reduces the number of indexes used during the training of the machine learning classifiers. It can be seen that the implementation of lemmatization increased the performance of both machine learning and hybrid approaches; meanwhile, the opposite observation has

been obtained by lexicon-based approach. This suggests that the lemmatization process is not needed or unnecessary in the pre-processing step of lexicon-based approaches, as the corpora has included different variations of the words containing the same meaning that will be used in calculating the sentiment score of the texts. In general, hybrid approaches outperformed both machine learning-based and lexicon-based approaches, as can be seen in Table 12. Furthermore, there is a tremendous increase in the performance by the hybrid approach against machine learning-based approach. In terms of the performance between different machine learning classifiers, it can be seen that Logistic Regression outperformed both Support Vector Machine and Naïve Bayes.

**Table 12.** The summary of the highest performing approaches using lemmatized texts for each dataset.

Dataset	Sentiment Analysis Approaches	Feature Extraction Technique	Logistic Regression	Support Vector Machine	Naïve Bayes
Earth Hour 2015 Corpus	Lexicon		Hu and Liu (90.5%)		
	Machine Learning	BoW	69.4%	74.9%	53.8%
		TF-IDF		71.5%	74.9%
	Hybrid	BoW		TextBlob (74.2%)	TextBlob (74.2%)
TF-IDF			Hu and Liu (72.5%)	VADER (74.2%)	TextBlob (64.8%)
Weather Sentiment	Lexicon		SentiStrength (58.8%)		
	Machine Learning	BoW	74.7%	70.4%	72.4%
		TF-IDF		74.3%	70.4%
	Hybrid	BoW		TextBlob (72.9%)	TextBlob (70.5%)
TF-IDF			TextBlob (73.1%)	TextBlob (70.5%)	TextBlob (64.7%)
Climate Change Sentiment	Lexicon		MPQA (54.0%)		
	Machine Learning	BoW	62.3%	60.4%	42.0%
		TF-IDF		62.6%	60.4%
	Hybrid	BoW		TextBlob (66.9%)	VADER (66.0%)
TF-IDF			TextBlob (70.4%)	TextBlob (76.6%)	WKWSCI (60.6%)
Combined Dataset	Lexicon		VADER (55.8%)		
	Machine Learning	BoW	70.0%	55.9%	62.6%
		TF-IDF		68.4%	55.9%
	Hybrid	BoW		TextBlob (74.7%)	TextBlob (61.8%)
TF-IDF			TextBlob (75.3%)	TextBlob (61.8%)	Hu and Liu (60.2%)

Furthermore, TextBlob dominated the majority of the hybrid approaches that were hybridized with either Logistic Regression or Support Vector Machine in terms of the best performance. One striking observation is the higher average F1-score of above 70% that was attained by Logistic Regression and Support Vector Machine in comparison with Naïve Bayes, which hovered around 60%. The Logistic Regression model which utilized TF-IDF outperformed those with BoW, however no significant improvement in the performance could be seen when it was used during the validation of Support Vector Machine and Naïve Bayes classifiers.

From the analysis of lexicon-based and hybrid approaches, it can be concluded that VADER and SentiWordNet achieved the highest correct classification rate for positive and negative tweets respectively. However, VADER and SentiWordNet tended to flip negative sentiment polarity to the opposite sign, which was positive, and vice versa. SentiStrength tended to misclassify negative tweets into neutral more frequently for lexicon-



based approaches. Meanwhile, this negative to neutral favoritism for hybrid approach includes the lexicons such as SentiStrength, Hu and Liu, and WKWSCl, that were combined with Logistic Regression.

Since the distinctions between negative with positive and negative with neutral are the most important factors to be clear about in investigating the sentiment of climate change and environment related tweets, thus it can be concluded that the lexicons that achieved this objective include TextBlob and MPQA for hybrid approaches and an addition of two other lexicons for lexicon-based approaches, which were Hu and Liu and WKWSCl. Since hybrid approach achieved the highest performance compared to the other methods, TextBlob with Logistic Regression as the classifier has been chosen as the most appropriate one to be used in climate change sentiment analysis tasks. Meanwhile, MPQA was not chosen because TextBlob had the advantage of achieving the highest performance in the majority of the observations for the hybrid approach.

## 5. Conclusions

This study was conducted with the aim to obtain the most effective method to be used in investigating climate change-related texts on social media platforms by performing a comparative evaluation of different sentiment analysis techniques, namely lexicon, machine learning, and hybrid approaches.

The hybrid between Logistic Regression and TextBlob with TF-IDF trained on lemmatized texts yielded an F1-score of 75.3% on the Combined Dataset. This concluded that hybrid approaches outperformed both lexicon and machine learning approaches. Furthermore, lemmatization is not recommended to be performed during the data preprocessing phase of lexicon approaches, as it can decrease the performance obtained. However, lemmatization has the effect of increasing the performance of both machine learning and hybrid approaches. This study also found that TF-IDF as the feature extraction technique outperformed BoW when it is used in Logistic Regression. Finally, no significant increase in performance can be seen when a bigger training size is used during the validation of machine learning classifiers. Future works will include investigating and comparing models produced using deep learning approaches toward this domain-specific sentiment on social media platforms by measuring the errors observed.

**Author Contributions:** Formal analysis, N.M.S.; Supervision, A.M.; Writing—original draft, N.M.S.; Writing—review & editing, A.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research is funded by the Ministry of Higher Education, Malaysia, and Institute of Research, Management and Innovation, Universiti Teknologi MARA under the Fundamental Research Grant Scheme (600-IRMI/FRGS 5/3 (370/2019).

**Data Availability Statement:** The datasets generated during and/or analyzed during the current study are available at the “Weather Sentiment” repository: <https://data.world/crowdfunder/weather-sentiment> (accessed on 1 June 2021). Meanwhile, for “Earth Hour 2015 Corpus” from the DecarboNet repository: <https://gate.ac.uk/projects/decarbonet/datasets.html> (accessed on 1 June 2021). and finally the third database used is Kaggle: <https://www.kaggle.com/joseguzman/climate-sentiment-in-twitter> (accessed on 1 June 2021).

**Acknowledgments:** We would like to express our deepest gratitude to the Ministry of Higher Education, Malaysia, and the Institute of Research, Management, and Innovation, Universiti Teknologi MARA for their continuous support.

**Conflicts of Interest:** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

1. World Meteorological Organization. *State of the Global Climate 2020*; World Meteorological Organization: Geneva, Switzerland, 2020.
2. Khan, M.Y.; Junejo, K.N. Exerting 2D-Space of Sentiment Lexicons with Machine Learning Techniques: A Hybrid Approach for Sentiment Analysis. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **2020**, *11*, 599–608. [[CrossRef](#)]

3. D’Aniello, G.; Gaeta, M.; La Rocca, I. KnowMIS-ABSA: An overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Artif. Intell. Rev.* **2022**, *1*–32. [[CrossRef](#)]
4. Xiang, N.; Wang, L.; Zhong, S.; Zheng, C.; Wang, B.; Qu, Q. How Does the World View China’s Carbon Policy? A Sentiment Analysis on Twitter Data. *Energies* **2021**, *14*, 7782. [[CrossRef](#)]
5. Agarwal, A.; Sharma, V.; Sikka, G.; Dhir, R. Opinion Mining of News Headlines using SentiWordNet. In Proceedings of the 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, India, 18–19 March 2016; pp. 1–5.
6. Sohagir, S.; Petty, N.; Wang, D. Financial Sentiment Lexicon Analysis. In Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 31 January–2 February 2018; pp. 286–289.
7. Jing, D.; Joyce, B. Sentiment analysis of tweets for the 2016 US presidential election. In Proceedings of the 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 3–5 November 2017.
8. Yadav, S.; Sarkar, M. Enhancing Sentiment Analysis Using Domain-Specific Lexicon: A Case Study on GST. In Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 19–22 September 2018; pp. 1109–1114.
9. Jung, J.; Petkanic, P.; Nan, D.; Kim, J.H. When a girl awakened the world: A user and social message analysis of Greta Thunberg. *Sustainability* **2020**, *12*, 2707. [[CrossRef](#)]
10. Khoo, C.S.G.; Johnkhan, S.B. WKWSC Sentiment Lexicon v1.1. 2017. Available online: <https://researchdata.ntu.edu.sg/dataset.xhtml?persistentId=doi:10.21979/N9/DWWEBV> (accessed on 1 May 2021).
11. Rustam, F.; Ashraf, I.; Mehmood, A.; Ullah, S.; Choi, G.S. Tweets Classification on the Base of Sentiments for US Airline Companies. *Entropy* **2019**, *21*, 1078. [[CrossRef](#)]
12. Gupta, I.; Joshi, N. Enhanced Twitter Sentiment Analysis Using Hybrid Approach and by Accounting Local Contextual Semantic. *J. Intell. Syst.* **2020**, *29*, 1611–1625. [[CrossRef](#)]
13. Mutanov, G.; Karyukin, V.; Mamykova, Z. Multi-Class Sentiment Analysis of Social Media Data with Machine Learning Algorithms. *Comput. Mater. Contin.* **2021**, *69*, 913–930. [[CrossRef](#)]
14. Zimbra, D.; Abbasi, A.; Zeng, D.; Chen, H. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans. Manag. Inf. Syst.* **2018**, *9*, 1–29. [[CrossRef](#)]
15. Alvi, M.B.; Mahoto, N.A.; Alvi, M.; Unar, M.A.; Shaikh, M.A. Hybrid Classification Model for Twitter Data—A Recursive Preprocessing Approach. In Proceedings of the 5th International Multi-Topic ICT Conference (IMTIC), Jamshoro, Pakistan, 25–27 April 2018; pp. 1–6.
16. Suhariyanto, A.; Firmanto; Sarno, R. Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet. In Proceedings of the 2018 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 21–22 September 2018; pp. 202–206.
17. Beigi, O.M.; Moattar, M.H. Automatic construction of domain-specific sentiment lexicon for supervised domain adaptation and sentiment classification. *Knowl.-Based Syst.* **2021**, *213*, 106423. [[CrossRef](#)]
18. Khoo, C.S.G.; Johnkhan, S.B. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *J. Inf. Sci.* **2018**, *44*, 491–511. [[CrossRef](#)]
19. Mahmood, A.; Kamaruddin, S.; Naser, R.; Nadzir, M. A combination of lexicon and machine learning approaches for sentiment analysis on Facebook. *J. Syst. Manag. Sci.* **2020**, *10*, 140–150.
20. Cai, M. Natural language processing for urban research: A systematic review. *Heliyon* **2021**, *7*, e06322. [[CrossRef](#)] [[PubMed](#)]
21. Guetterman, T.C.; Chang, T.; De Jonckheere, M.; Basu, T.; Scruggs, E.; Vydiswaran, V.V. Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological Study. *J. Med. Internet Res.* **2018**, *20*, e9702. [[CrossRef](#)] [[PubMed](#)]
22. Casey, A.; Davidson, E.; Poon, M.; Dong, H.; Duma, D.; Grivas, A.; Grover, C.; Suárez-Paniagua, V.; Tobin, R.; Whiteley, W.; et al. A systematic review of natural language processing applied to radiology reports. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 179. [[CrossRef](#)]
23. Khairi, N.I.; Mohamed, A.; Yusof, N.N. Feature Selection Methods in Sentiment Analysis: A Review. In Proceedings of the the 3rd International Conference on Networking, Information Systems & Security, Marrakech, Morocco, 31 March 2020–2 April 2020; pp. 1–7. [[CrossRef](#)]
24. Keshavarz, H.; Abadeh, M.S. Accurate frequency-based lexicon generation for opinion mining. *J. Intell. Fuzzy Syst.* **2017**, *33*, 2223–2234. [[CrossRef](#)]
25. Ahmad, M.; Aftab, S.; Muhammad, S.S.; Waheed, U. Tools and Techniques for Lexicon Driven Sentiment Analysis: A Review. *Int. J. Multidiscip. Sci. Eng.* **2017**, *8*, 17–23.
26. Ramanathan, V.; Meyyappan, T. Twitter Text Mining for Sentiment Analysis on People’s Feedback about Oman Tourism. In Proceedings of the 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 15–16 January 2019; pp. 1–5.
27. Nahar, K. Social Media Sentiment Analysis: The Hajj Tweets Case Study. *J. Comput. Sci.* **2021**, *17*, 265–274.
28. Machuca, C.R.; Gallardo, C.; Toasa, R.M. Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach. *J. Phys. Conf. Ser.* **2021**, *1828*, 012104. [[CrossRef](#)]
29. Rajput, A.E. Natural Language Processing, Sentiment Analysis and Clinical Analytics. *arXiv* **2019**, arXiv:1902.00679.
30. Alsaeedi, A.; Khan, M.Z. A Study on Sentiment Analysis Techniques of Twitter Data. *(IJACSA) Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 361–374. [[CrossRef](#)]

31. Bonta, V.; Kumaresh, N.; Janardhan, N. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian J. Comput. Sci. Technol.* **2019**, *8*, 1–6. [CrossRef]
32. Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment Strength Detection in Short Informal Text. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2544–2558. [CrossRef]
33. Riloff, E.; Wiebe, J. Learning Extraction Patterns for Subjective Expressions, 2003. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 11–12 July 2003.
34. Nasim, Z.; Rajput, Q.; Haider, S. Sentiment Analysis of Student Feedback Using Machine Learning and Lexicon Based Approaches. In Proceedings of the 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, Malaysia, 16–17 July 2017; pp. 1–6.
35. Ligthart, A.; Catal, C.; Tekinerdogan, B. Systematic reviews in sentiment analysis: A tertiary study. *Artif. Intell. Rev.* **2021**, *54*, 4997–5053. [CrossRef]
36. Bhavitha, B.K.; Rodrigues, A.P.; Chiplunkar, N.N. Comparative Study of Machine Learning Techniques in Sentimental Analysis. In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 10–11 March 2017; pp. 216–221. [CrossRef]
37. Amin, S.; Uddin, M.I.; Al-Baity, H.H.; Zeb, M.A.; Khan, M.A. Machine Learning Approach for COVID-19 Detection on Twitter. *Comput. Mater. Contin.* **2021**, *68*, 2231–2247. [CrossRef]
38. Das, D.D.; Sharma, S.; Natani, S.; Khare, N.; Singh, B. Sentiment Analysis for Airline Twitter data. *IOP Conf. Ser. Mater. Sci. Eng.* **2017**, *263*, 042067.
39. Lalji, T.K.; Deshmukh, S.N. Twitter Sentiment Analysis Using Hybrid Approach. *Int. Res. J. Eng. Technol. (IRJET)* **2016**, *3*, 2887–2890.
40. Rajeswari, A.M.; Mahalakshmi, M.; Nithyashree, R.; Nalini, G. Sentiment Anaysis for Predicting Customer Reviews using a Hybrid Approach. In Proceedings of the 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), Cochin, India, 2–4 July 2020.
41. Angiani, G.; Ferrari, L.; Fontanini, T.; Fornacciarri, P.; Iotti, E.; Magliani, F.; Manicardi, S. *A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter*; KDWeb: London, UK, 2016.
42. Krouska, A.; Troussas, C.; Virvou, M. The effect of preprocessing techniques on Twitter sentiment analysis. In Proceedings of the 2016 7th International Conference on Information Intelligence, Systems & Applications (IISA), Chalkidiki, Greece, 13–15 July 2016; pp. 1–5.
43. Pradha, S.; Halgamuge, M.N.; Vinh, N.T.Q. Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data. In Proceedings of the 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Da Nang, Vietnam, 24–26 October 2019; pp. 1–8.
44. Zhao, J.; Gui, X. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access* **2017**, *5*, 2870–2879.
45. Mutinda, J.; Mwangi, W.; Okeyo, G. Lexicon-pointed hybrid N-gram Features Extraction Model (LeNFEM) for sentence level sentiment analysis. *Eng. Rep.* **2021**, *3*, e12374. [CrossRef]
46. Rustam, F.; Khalid, M.; Aslam, W.; Rupapara, V.; Mehmood, A.; Choi, G.S. A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis. *PLoS ONE* **2021**, *16*, e0245909. [CrossRef]
47. Thelwall, M. SentiStrength. 2010. Available online: <http://sentistrength.wlv.ac.uk> (accessed on 1 March 2021).
48. Crowdfunder. data.world, 22 November 2016. Available online: <https://data.world/crowdfunder/weather-sentiment> (accessed on 1 June 2021).
49. Maynard, D.; Bontcheva, K. GATE, May 2016. Available online: <https://gate.ac.uk/projects/decarbonet/datasets.html> (accessed on 1 June 2021).
50. Maynard, D.; Bontcheva, K. Challenges of evaluating sentiment analysis tools on social media. In Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016; pp. 1142–1148.
51. Guzman, J. Kaggle, 24 December 2020. Available online: <https://www.kaggle.com/joseguzman/climate-sentiment-in-twitter> (accessed on 1 June 2021).
52. Elbagir, S.; Yang, J. Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. In Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong, China, 13–15 March 2019.
53. Symeonidis, S.; Effrosynidis, D.; Arampatzis, A. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Syst. Appl.* **2018**, *110*, 298–310. [CrossRef]
54. Balakrishnan, V.; Ethel, L.-Y. Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lect. Notes Softw. Eng.* **2014**, *2*, 262–267. [CrossRef]
55. Wilson, T.; Wiebe, J.; Hoffmann, P. Subjectivity Lexicon. 2005. Available online: [http://mpqa.cs.pitt.edu/lexicons/subj\\_lexicon/](http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/) (accessed on 1 May 2021).
56. Koncz, P.; Paralic, J. An approach to feature selection for sentiment analysis. In Proceedings of the 2011 15th IEEE International Conference on Intelligent Engineering Systems, Poprad, Slovakia, 23–25 June 2011; pp. 357–362.
57. Zhao, R.; Mao, K. Fuzzy Bag-of-Words Model for Document Representation. *J. Latex Cl. Files* **2015**, *14*. [CrossRef]

58. Garg, S. Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning. In Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 28–29 January 2021; pp. 175–181.
59. Ramasamy, L.K.; Kadry, S.; Nam, Y.; Meqdad, M.N. Performance analysis of sentiments in Twitter dataset using SVM models. *Int. J. Electr. Comput. Eng.* **2021**, *11*, 2275–2284. [[CrossRef](#)]
60. Roshani, M.; Sattari, M.A.; Ali, P.J.M.; Roshani, G.H.; Nazemi, B.; Corniani, E.; Nazemi, E. Application of GMDH neural network technique to improve measuring precision of a simplified photon attenuation based two-phase flowmeter. *Flow Meas. Instrum.* **2020**, *75*, 101804. [[CrossRef](#)]
61. Charandabi, S.E.; Kamyar, K. Using A Feed Forward Neural Network Algorithm to Predict Prices of Multiple Cryptocurrencies. *Eur. J. Bus. Manag. Res.* **2021**, *6*, 15–19. [[CrossRef](#)]
62. Dizadji, M.R.; Yousefi-Koma, A.; Gharehnazifam, Z. 3-Axis Attitude Control of Satellite using Adaptive Direct Fuzzy Controller. In Proceedings of the 2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM), Tehran, Iran, 23–25 October 2018; pp. 1–5.
63. Dizaji, M.R.; Yazdi, M.R.H.; Shirzi, M.A.; Gharehnazifam, Z. Fuzzy supervisory assisted impedance control to reduce collision impact. In Proceedings of the 2014 Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM), Tehran, Iran, 15–17 October 2014; pp. 858–863.
64. Abdelwahab, O.; Bahgat, M.; Lowrance, C.J.; Elmaghraby, A. Effect of training set size on SVM and Naive Bayes for Twitter sentiment analysis. In Proceedings of the 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Abu Dhabi, United Arab Emirates, 7–10 December 2015; pp. 46–51.