School of Electronic Engineering and Computer Science

Programme of study:
BSc Computer Science

Project Title:
"Accuracy in Sentiment Analysis: An ensemble approach"

Supervisor:
Dr. Miles Hansard

Student Name:
Yanitsa Stancheva

Final Year
Undergraduate Project 2022/23

Date: 09/05/2023

Contents

List of Figures

List of Tables

Abstract

As the world becomes ever more enmeshed with social media, these platforms represent an excellent source of information around what people think and feel about. One increasingly used method to tap into this information is Sentiment Analysis. The accuracy of such a method is of significant importance if it is to be useful in the wider world. Furthermore, the sort of language used in social media is often figurative and harder to identify. To address this problem of accuracy the current study utilises three supervised and one unsupervised method to classify two different datasets, one revolving around Climate Change the other focusing on figurative language. Having ascertained the best two models for classification, an ensemble model was used to demonstrate greater accuracy than any one classifier. The ensemble method performed as well as the best classifier for the Climate Change dataset. While the accuracy for Sarcasm and Irony detection was improved. Such results add to the growing body of evidence that ensemble methods may represent a better way to address the problem of accuracy in Sentiment Analysis.

Chapter 1: **Introduction**

1.1 Background

The world's use of social media has dramatically increased over the past decade. In part due to the accessibility of the technology. Smart phones, tablets and other technologies are making it ever easier for users to access social media services. These services are significant in that they profoundly impact how we build relationships, access information and engage with others on matters that are of importance. As a result, opinions that are expressed on social media can play a significant role in shaping public discourse and influencing large political issues. Social media platforms provide an avenue for individuals to express their opinions and engage with others who have similar or differing views. This can lead to the formation of online communities that may mobilize around specific issues, leading to larger movements or even protests. Furthermore, politicians and business may use the data expressed on social media to gauge the sentiment held by the public on issues relating to them. The impact of social media on the publics lives has been profound. However, the abuse of that data is also widespread. The Cambridge-Analytica scandal was one such example of data from social media being used for nefarious means (New York Times, 2018).

One area in which social media has been significantly associated is the sentiment held by the public surrounding global warming. Regardless of the individual opinions held by the public on the issue it has nonetheless been a hot topic. The effect of global warming can be felt globally with sea levels rising and more severe local weather effects. For example, countries such as France and the UK have experienced severe droughts and record temperatures in 2022. As more greenhouse gases accumulate in the atmosphere temperatures are only predicted to rise. NASA has predicted that the world will experience more droughts, severe weather, and a continued rise in sea levels well into the next century (NASA, 2022). Given the importance of these issues, global warming has become one of the most discussed topics within the public and political spheres of discourse. This was most evident when in 2015, 196 countries all collectively agreed to take action to fight global warming, when they each signed the Paris Agreement at the UN Climate Change Conference (UNFCC, 2023).

As social media has become more prevalent in people's daily lives it has also become an ever-increasing source of information. Data has shown that in the UK 49% of the public use social media as a source of news

(OFCOM, 2022). Furthermore, in September 2022, the Office of National Statistics reported that three in four adults, aged 16 and over are feeling worried about global warming (ONS, 2022). The sentiments and opinions expressed on social media can therefore influence political decision-making, as politicians may adjust their policies or messaging based on the views of their constituents and the wider public.

The accurate monitoring of opinions and sentiment on social media are important for policymakers because they can provide insights into what the public attitude is towards important issues. Traditional approaches such as surveys and focus groups are good examples of methods used to gain an insight into the wider public sentiment and opinions on important issues. Such methods are, however, usually expensive to conduct and may have some delay in returning results. With the exponentially increasing ubiquity and popularity of various social media platforms, like Twitter or Facebook, massive amounts of commentary data have now become more accessible. An official statistic published by Insider Intelligence in 2022, showed that there are 353.9 million active users on Twitter (Insider Intelligence, 2022).

An increasingly prominent method for gathering and analysing data from a social media platform like Twitter, has been Sentiment Analysis. Sentiment Analysis refers to “the general method to extract subjectivity and polarity from text” (Taboada et al, 2011). This subjectivity contains within it both opinions and sentiments held by an individual. It is the latter that sentiment analysis unsurprisingly addresses. In other words, sentiment analysis provides a means with which to extract what people feel about something given what they have expressed.

1.2 Problem Statement

Social media is a significant expression of people's sentiments surrounding issues or people. There are various means with which to classify the sentiments expressed on social media. One problem with sentiment analysis is the accuracy of the classifier used especially as there are a wide variety of methods with which to classify data. Furthermore, given that language is rich with nuance, sentiment analysis needs to be able to detect such subtleties. Sarcasm and irony are examples of how language is used differently to convey emotion. Sarcasm can be both positive and negative (Example: "Yeah, you have been sooo helpful" or "That is exactly what I needed today") and thus makes it challenging to classify and produce accurate results. Irony on the other hand might appear as "What a beautiful day" when in fact the weather outside was appalling. This example demonstrates how the classifier would need to detect that what is expressed is in fact not what is actually felt.

1.3 Aims and Objectives

This project aims at building a model for obtaining sentiment towards climate change using content posted in a micro-blogging format. The data will be pre-processed and classified using Naïve Bayes, Random Forest, Support Vector Machine, and an unsupervised lexicon-based approach. These models were chosen as they represent a wide variety of methods with which to classify data. Naïve Bayes represents a probabilistic method to approach sentiment classification; Random Forest provides insight into the accuracy of ensemble decision trees in classifying sentiment, while Support Vector Machine provides a statistical underpinning to classifying sentiment. Finally, the lexicon-based approach represents the accuracy with which a model can classify sentiment without any previously-held-knowledge. This differs from the other three supervised models as it is entirely unsupervised.

Furthermore, this project aims to investigate the differences between the classifiers with respect to their ability to detect sarcasm and irony. These two linguistic phenomena are some of the many that can confound models when they attempt to classify sentiment. This is of particular importance given that these linguistic phenomena tend to contain significant emotional content. However, this is even more difficult to detect given that often sarcasm or irony can be ambiguous as to its sentiment.

Furthermore, the use of two classifiers in an ensemble method will be performed. The aim of which is to determine if different approaches to the sentiment analysis problem can be combined. The intention being that the ensemble method will produce even greater accuracy.

1.4 Report Structure

This report presents the steps taken to fulfil the aims and objectives presented in this chapter. The second chapter will provide a review of previous research and current approaches available in the literature. In the third chapter, the proposed workflow and the applied methods will be presented in detail. Having developed and implemented the proposed methods, all will be evaluated, and the results will be presented in the fourth chapter. The fifth chapter includes a discussion of the results detailing their performance, evaluates each approach, and their pros and cons. The report ends with the sixth chapter which summarises the findings, discusses the overall achievements, challenges, and provides possible avenues of expanding this research.

Chapter 2: **Literature Review**

2.1 Overview

Sentiment analysis has not been around for a particularly long time. Some of the earliest examples involved classifying the sentiment held within reviews. For example, Pang et al. (2002) investigated how it could be possible to classify the reviews of movies in terms of whether the sentiment was positive or negative. Similarly, Turney (2002) also demonstrated a learning algorithm that was capable of classifying movie reviews in terms of whether they were "recommended" (positive) or "not recommended" (negative). Such examples were attempting to demonstrate a machine learning capacity for classifying the sentiment expressed. Prior to these studies, much of the work that had been performed around sentiment classification had utilised some form of knowledge-based classification. Pang and their colleagues endeavoured to use completely prior-knowledge-free supervised machine learning methods (2002). Alternatively, Turney made use of an unsupervised model.

As society has moved ever further towards the digital sphere for expression, sentiment analysis has increasingly been used as a tool to collect information around the feelings people hold around issues or objects. Such usage has created a distinction between an opinion and a sentiment. According to the Oxford dictionary an opinion is defined as, "your feelings or thoughts about somebody/something, rather than a fact", while a sentiment is defined as, "a feeling or an opinion, especially one based on emotions". These two definitions naturally feature a great deal of overlap. As Pozzi and their colleagues (2016) describe, "the definitions indicate that an opinion is more of a person's concrete view about something, whereas a sentiment is more of a feeling". The former therefore representing a held assessment of someone or something, but as the dictionary definition emphasises that is not necessarily a factually held assessment. Sentiment conversely represents feelings or opinions, but critically is based on emotions.

Bing Lui (2020) provides very clear examples for the difference between an opinion and a sentiment. If, for example we take the phrase, "I am concerned about the current state of the economy" we can see that such a phrase encompasses clear sentiment as it uses the word "concern" to convey a feeling about the economy. Conversely if we take the phrase, "I think the economy is not doing well" we can see that it conveys an assessment about the economy, whether that is factual or not. Furthermore, Lui illustrates that the way in which we respond to these statements also

illuminates if they are a sentiment or an opinion. So, for the first example, people would likely respond with some sort of emotional response to the concern being expressed. For example, with a statement conveying that they share the concern about the economy. With the latter example, the expectation is that the person being spoken to would respond with some sort of agreement or disagreement rather than an expression of feeling about the topic of discussion. (Ibid, 2020)

Sentiment analysis utilises a varied spectrum of machine learning tools. These models are either supervised and therefore rely on prior held knowledge, or unsupervised and as such are not based on any prior held knowledge. One such notable supervised model is the Naïve Bayes Classifier. This model makes use of probabilities to classify sentiment. It is a somewhat straightforward classifier that requires the data to be pre-processed. However, a drawback of such a model is that it relies on the assumption that each word from the text to be classified is independent from every other word. This, as Gamallo and his colleagues (2012) have noted, creates an inevitable restriction, as regardless of the process undertaken to pre-process the data, words inevitably have "syntactic and semantic dependencies". In the study reported by Gamallo and his colleagues, the Naïve Bayes Classifier reported an accuracy of 67%. In a more recent study (2021), the authors made a comparison of five different classifiers including the Naïve Bayes Classifier as well as: SVM, AdaBoost classifier, K-nearest neighbour, and a decision tree classifier. The Naïve Bayes Classifier outperformed all the other classifiers with an accuracy of 61.2%.

An alternative model to address classification is the Random Forest Classifier. This classifier is also within the group of supervised classifiers that makes use of ensembles of decision trees to classify sentiments. The group of decision trees are combined to form a final sentiment output. Unlike Naïve Bayes there is no assumption of conditional independence or probabilities whatsoever. Furthermore, the use of multiple decision trees enhances the accuracy by reducing the noise present in a single decision tree (Zahoor & Rohilla, 2020). In one study that examined utilising both Random Forest and Support Vector Machine, the authors found that the Random Forest correctly classified the data with an accuracy of 81% while the Support Vector Machine achieved an accuracy of 82.4%. Interestingly the authors combined the 2 models into a hybrid classifier and achieved an improved accuracy (Al Amrani et al., 2018). In another study the authors demonstrated accuracy with Random Forest Classifier at 85% accuracy compared to Support Vector Machine achieving an accuracy of 81.5%. (Saifullah et al., 2021) The study made use of social media data surrounding Covid 19 and the anxiety that was being expressed online.

Classification can also rest on statistical means and forms the basis for the Support Vector Machine method. This model of classification was initially conceived of by Vladimir Vapnik in the 1990's and was initially predicated on classifying data in to only two classes (Al Amrani et al., 2018). It has since developed further and can now be used to classify data into more than simply two classes. It can now be used to address continuous outcomes allowing it to be useful across a far wider spectrum of data and classification tasks. For example, the capacity to address continuous outcomes allows the model to be used for regressions as well as classifications (Guenther & Schonlau, 2016) The Support Vector Machine model is concerned with separating data across what it terms a 'hyper-plane' to create a linear non-probabilistic outcome. The model generates the 'hyper-plane', as it represents the best line capable of separating the data into classes. Naturally this is best used for linear data however it has strong efficacy even when the data is non-linear. (Al Amrani et al., 2018) Zhoor and their colleague investigated a comparison between Support Vector Machine, Naïve Bayes Classifier, Random Forest and also Long Short Term Memory Networks. When comparing the first three classifiers the authors found that all of them demonstrated accuracies in the region of 88-98%. However, there was disparity between which classifier performed the best given one of the five datasets. For example, when using datasets comprised of tweets, the Naïve Bayes Classifier performed the best. Whereas, when using reviews of a movie the Support Vector Machine classifier performed the best (Zhoor & Rohilla, 2020). This raises questions around the linguistic differences between tweets and say movie reviews and the model's capacity to classify them correctly.

As has already previously mentioned there are of course unsupervised methods of classification. All the three models for classification that have been described thus far represent supervised methods, despite each having a different underpinning. The lexicographic approach is an example of an unsupervised approach. In this method the data is parsed by making use of a lexicon of words with which it can make comparisons. So, for example, the model can identify whether the data provided is positive, negative, or neutral by referring to dictionaries. These can be manually created, or the model can make use of automatically generated dictionaries. Some examples of commonly used lexicons are General Inquirer or Linguistic enquiry (Mehto & Indras, 2016). One of the earliest examples of the lexicon-based approach was demonstrated by Turney in 2002. As was touched on earlier this method was entirely unsupervised and demonstrated a capacity to classify movie reviews as either "recommended" or "not recommended" without any prior-held knowledge provided to the model. It is this capacity which separates unsupervised models from supervised models. Within the discourse surrounding lexicon-based approaches, Hutto and Gilbert (2014)

demonstrated a robust lexicon that can be used to classify data by sentiment. In their study they compared various lexicons against theirs, as well as against several supervised machine learning models. Two of the supervised models have been touched on already, namely, Naïve Bayes and Support Machine Vector. Hutto and Gilbert demonstrated that their lexicon VADER is more than capable of classifying the sentiment held within social media text. Their results demonstrated that it is even capable of outperforming independent human classifiers, who outperformed any of the other supervised models. It is worth noting that they also used reviews and editorial datasets, the results of which showed VADER was often better than supervised models but not better than human classifiers.

As powerful a tool as Sentiment Analysis is, it is inevitably confounded by many linguistic practices that are rife within language. For example, someone may express something via a tweet in which they do not indicate how they feel about something that is already emotional. Does the model then classify them as neutral, or as positive or negative. Similarly, the speaker may in fact direct different sentiment to different targets within the same text. How then can the model determine the sentiment that is being expressed (Mohammad, 2017). Other examples include the detection of idioms and emoji. Facets of language that do not follow the usual rules of interpretation. One area of language that can trip up sentiment classification is of course figurative language. Language, wherein the meaning being expressed by the speaker is in fact different to the literal interpretation of what was expressed. Sarcasm and irony are two significant examples of figurative language, and they are prevalent within the sphere of social media. Dealing with these types of figurative language present a greater problem when classifying social media texts. Farias and Rosso (2017) point out that social media texts are "informal and use ill-formed language". This is compounded by the fact that often such communications have errors, abbreviations, and colloquialisms. In the specific case of Twitter, the limit of 140 characters drives the need to perhaps express things without the usual semantic and syntactic rules. As a result, detecting sarcasm and irony in social media texts presents a significant challenge when classifying sentiment.

2.2 Sentiment Analysis Workflow

The typical workflow in sentiment analysis involves the following steps:

1. Data Collection: The first step is to collect data that will be used for sentiment analysis. This can involve scraping social media platforms, review websites, or any other source of text data.

2. Data Pre-processing: Once the data has been collected, it needs to be pre-processed to prepare it for analysis. This typically involves removing “stopwords”, stemming or lemmatizing words, and converting all text to a consistent case.

3. Feature Extraction: The next step is to extract features from the pre-processed text that will be used to train a machine learning model. Common features include bag-of-words, n-grams, or word embeddings.

4. Model Training: With the features extracted, the next step is to train a machine learning model on a labelled dataset. There are many different algorithms that can be used for sentiment analysis, including Naive Bayes, Support Vector Machines (SVM), and Recurrent Neural Networks (RNN). For the purposes of this project Naïve Bayes, Random Forest Classifier, Linear Support Vector Classifier (SVC) and Lexicographic approach will be used.

5. Model Evaluation: Once the model is trained, it needs to be evaluated to determine its performance on unseen data. This typically involves splitting the labelled dataset into training and test sets, and using metrics such as accuracy, precision, recall, and F-score to evaluate the model.

6. Model Deployment: If the model meets the desired performance criteria, it can be deployed to make predictions on new, unlabelled data.

7. Model Monitoring: Finally, it's important to monitor the performance of the deployed model over time to ensure that it continues to perform well and remains accurate as new data becomes available.

A visual representation of the Sentiment Analysis work flow can be seen in Figure 1

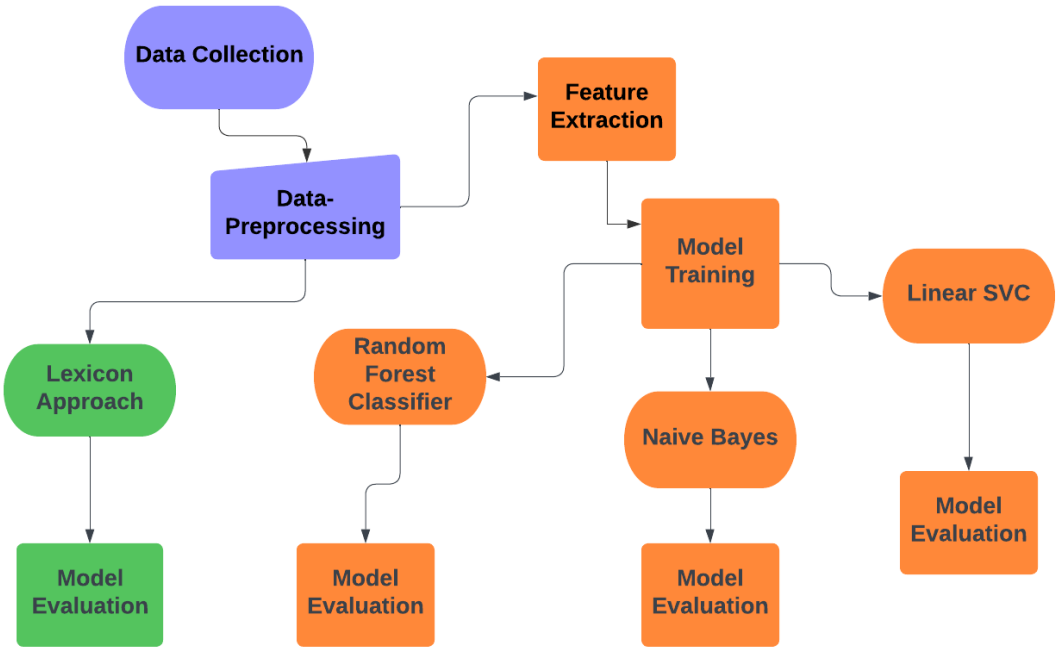

Figure 1 Sentiment Analysis Workflow

2.3 Classification Methods

2.3.1 Supervised

This method uses labelled data in order to train the model and it can be classified into two categories: classification and regression. In classification the aim to assign the input data into categories, based on the number of classes they are either binary (when there are only two classes) or multiclass classification, when the goal is to predict the class label for each input. According to Kelleher and Tierney (2018), the most common machine learning is the supervised one, making it a preferred choice in a number of domains such as natural language processing (NLP), speech or image recognition. There are numerous examples of supervised methods, as three supervised models will be used in this body of work, they will be discussed further in Chapter 3.

2.3.2 Unsupervised

Unsupervised methods in sentiment analysis refer to techniques that do not rely on labelled data or a pre-defined set of categories. These methods instead use statistical and linguistic models to identify patterns in text that indicate positive, negative, or neutral sentiment. Typical unsupervised techniques for sentiment analysis include:

1. Lexicon-based approach: This method uses dictionaries (or lexicons) that associate words with either positive, negative, or neutral scores to the sentiment, it then calculates it as the sum or average of the sentiment scores of the words in the text.
2. Machine learning clustering algorithms: Unsupervised algorithm that groups similar datapoints together.
3. Latent Dirichlet Allocation (LDA): LDA is a topic modelling algorithm that identifies latent topics in a corpus of documents. Sentiment can be inferred by examining the sentiment of the words associated with each topic.
4. Deep Learning: Models such as Auto-Encoders and Variational Auto-Encoders can learn the representations of the text without any labels. These representations can be used for downstream tasks such as sentiment analysis.

2.3.3 Semi-Supervised

Semi-supervised methods in sentiment analysis involve using a combination of labelled and unlabelled data to train a sentiment classifier. The idea is that while labelled data is expensive and time-consuming to obtain, unlabelled data is abundant and relatively easy to acquire. Some common semi-supervised methods in sentiment analysis include:

1. **Self-training:** This method involves training a classifier on a small set of labelled data, and then using it to classify the remaining unlabelled data. The most confident predictions are added to the labelled dataset, and the process is repeated iteratively.
2. **Co-training:** This method involves training two classifiers on different sets of features or views of the data and using them to label each other's unlabelled data. This allows for better generalization and can improve performance when there is a limited amount of labelled data available.

3. **Active learning:** This method involves selecting the most informative examples from the unlabelled data and asking a human annotator to label them. These examples are then used to update the classifier, and the process is repeated iteratively.

2.6 Current Challenges

Sentiment analysis is a challenging task that involves analysing and understanding the emotions, opinions, and attitudes expressed in textual data. Some of the challenges in sentiment analysis include:

1. **Contextual understanding:** Sentiment analysis models need to understand the context in which the text was written to determine the sentiment more accurately. For example, in the sentence "I love this movie" the context could be positive if it was written by someone with positive experience, but negative if it was written by someone who dislikes the film.
2. **Ambiguity:** Ambiguous phrases or words that can have different meanings depending on the context of the text. This can lead to sentiment misclassification.
3. **Irony and sarcasm:** Understanding sarcasm is a challenge for Sentiment Analysis as it involves emotional factors such a facial expression, change in voice intonation and context.
4. **Negation:** The presence of negation words such as "not" and "never" can invert the polarity of sentiment, thus making it difficult for models to accurately classify the sentiment.
5. **Domain-specific language:** Sentiment analysis models trained on generic datasets may not perform well when applied to domain-specific language. Technical jargon used in healthcare or finance is an example of that.
6. **Data imbalance:** Sentiment analysis models can be biased towards the class that holds majority in the dataset. This can lead to inaccurate predictions for minority classes.
7. **Multilingualism:** Different languages can be a challenge to Sentiment Analysis, due to grammar, syntax or cultural factor.

Chapter 3: **Methodology**

At the outset, the current project aims to use machine learning to train a large, labelled dataset that is therefore capable of classifying the sentiment in the datasets. As such, this is known as a classification problem. To address the classification problem, both supervised and unsupervised methods will be utilised. The supervised approach will make use of several different classification algorithms. These classifiers will be: Naïve Bayes Classifier, Random Forest, and Linear Support Vector Classifier. These classifiers are being used as they reflect a varied basis from which the classification problem is to be addressed. Finally, a Lexicon based approach will be used to tackle the classification problem within an unsupervised framework. This will mean that the first three classifiers will classify the datasets with the addition of prior-held-knowledge. The lexicon-based approach will be entirely unsupervised and therefore rely in no way on any prior-held-knowledge. After testing all the above classifiers, Linear SVC and Naïve Bayes will be combined into an ensemble method.

Finally, all programming and calculations were carried out using Python 3.10, ScikitLearn 1.2.2, NLTK 3.8.1, and Pandas 1.5.3. Other Python libraries and dependencies can be found in the ‘requirements.txt’ file in the project folder submitted as a part of this report.

3.1 Datasets Used

3.1.1 Climate Change Tweets

For the purposes of this project, a twitter dataset containing 34.7k tweets wherein users are expressing their sentiment towards climate change is used (Kaggle). The overall structure of this dataset includes the message of the user, the tweet ID and the marked sentiment of each message ranging from -1 to 1 (-1 for negative, 0 for neutral and 1 for positive). The distribution of the sentiments is shown in Figure 2.

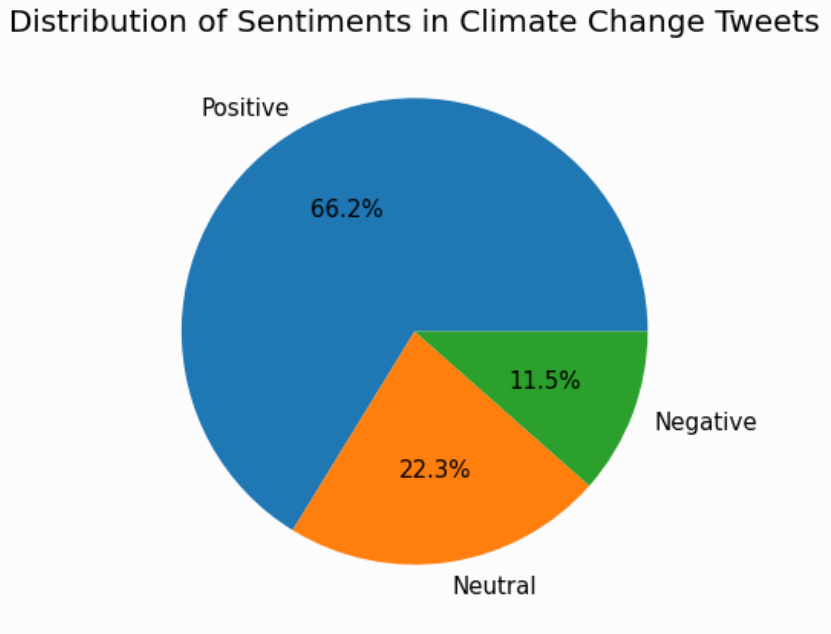

Figure 2 - Distribution of Sentiments in Climate Change Tweet Dataset

The labelling of this twitter dataset has been carried out by 3 different persons with the final label being based on majority consensus. Positive sentiment here is defined as pro-environmentalist stance. For example, that climate change is real, its impact should be mitigated, and awareness raising action should be taken, are examples of a Positive sentiment in this context. Negative sentiments generally revolve around downplaying or outright rejecting the scientific consensus on climate change.

3.1.2 Irony and Sarcasm Tweets

The second Twitter dataset contains 66254 tweets on a mixture of tweets containing tweets with sarcasm, irony, and regular (Klinger, 2016). These tweets have been annotated by the users using hashtags crawled from Twitter between July and September of 2015. Each tweet is labelled as "ironic", "sarcasm", and "regular".

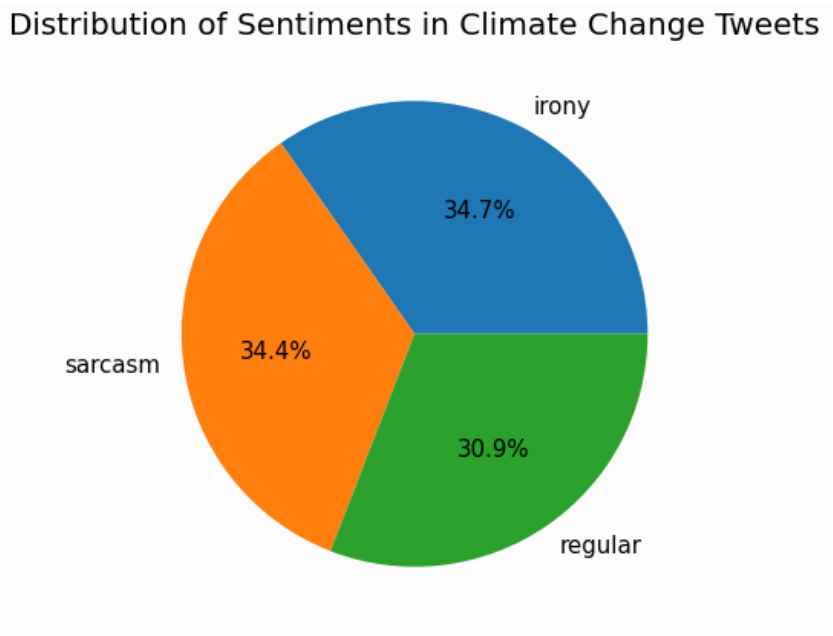

Figure 3 - Distribution of Sentiments in Climate Change Tweet Dataset

3.2 Data Cleaning and Preparation

Before this data is "fed" into the model, it needs to be prepared by following the following steps:

1. **URL links:** URL links and hashtags were replaced with a word placeholder and later removed in the tokenization process.
2. **Stop Words:** removing commonly used words in personal pronouns such as "a", "the", "we", "he", "she". These are pronouns from everyday language and do not carry any weight over the analysis but parsing them will increase the workload on the system. In order to achieve this, the English 'stopwords' module of NLTK library in Python was utilized.
3. **Capital case letters:** All capital letters were replaced with lowercase. Capital letters could present a potential issue. An example of that is

"US" such as in the United States and "us" which is represented as "we".

4. **Stemming and lemmatization:** Trimming the end of words that have the same meaning and reducing them to their common base (Stanford University, 2008). Example of such word is "change", "changing", "changes", "changed". The WordNetLemmatizer from NLTK library in Python was used to achieve this.
5. **Abbreviation:** Abbreviation is a method of replacing slang with their respective words. Often in messaging platforms or social media there is a big presence of abbreviations, however the model will not be able to recognise them. This study uses the ChatSlang dictionary that contains over 400 of the most commonly used abbreviations, such as "lol" (laugh out loud), "ttyl" (talk to you later).
6. **Tokenization:** Tokenization is a method of breaking down a sentence into smaller chunks of individual words which are called "tokens".

   By performing this process, it helps machines understand human language at a more granular level.
   An example of tokenization is given the sentence "I feel like today will be a wonderful day" to output the result as "I", "feel", "like". "today", "will", "be", "a", "wonderful", "day".

3.3 Lexicographic Model

The lexicon-based approach is a commonly used method in sentiment analysis, which involves the use of pre-defined lists of words or phrases, known as lexicons. This helps to determine the polarity (positive, negative, or neutral) of a given text. In the lexicographic model, each word in the text is matched against the words in the lexicon, and the resulting score is used to determine the overall sentiment of the text (Hutto & Gilbert, 2014).

The lexicographic model does not require extensive training data or high computational power, which is one of its advantages. However its limitations are inability to capture effectively the complexity of the human language, such as idiomatic expressions or sarcasm.

It is a simple, yet effective method in analysing the sentiment that has been expressed in a text. The current project uses one of the widely used lexicons developed by Finn Årup Nielsen in 2011 called AFINN. This

lexicon consists of 3,300 words along with phrases in English, each one of them come with a score ranging from -5 to +5, with -5 indicating a very negative sentiment whereas +5 indicates very strong positive sentiment.

3.4 Naïve Bayes Classifier

The Naïve Bayes Classifier utilises a supervised machine learning procedure, whereby the model learns and identifies patterns based on labelled data provided to it. The term "Naïve" refers to the basic assumption of there being strong independence between the features of the model. This so called "conditional independence" assumes that one feature is independent from another feature. In other words, the probability of a tweet belonging to a certain sentiment class is predicated on assuming that each word used in a tweet is conditionally independent from every other word that is present in the tweet.

In terms of this research, the algorithm is concerned with calculating the probability of each sentiment class given the data that is provided to it. This is achieved by calculating the product of the conditional probabilities of each word given its sentiment class. This is then multiplied by the result of the prior probability of that sentiment class. As a result, this can therefore be expressed formulaically as:

$$P(C|X) = \frac{P(C) * P(X|C)}{P(X)}$$

From the given formula, C refers to the sentiment class (Positive Negative, or Neutral) whereas X refers to the input data.

$P(C)$ refers to the prior probability of the of the class. $P(X|C)$ is the conditional probability from the input data, given the class. $P(X)$ refers to the probability of the input data (tweet for climate change). The formula could be simplified as follows:

$$P(X|C) = P(x1|C) * P(x2|C) \dots P(xn|C)$$

where $x1, x2, \dots, xn$ are the features of the input data (M. K. Leung, 2007).

The Multinomial Bayes model works with text classification problems where the input data consists of word counts(word frequencies in text data).

It is called "multinomial", because it models the distribution of word counts across multiple classes. It is easy to implement and is one of the most effective classification algorithms (Shajahan, 2022).

3.5 Random Forest Classifier

Random Forest Classifier much like Naïve Bayes classifier is a supervised machine learning technique. However, unlike the former technique Random Forest Classifier does not rely on probabilities. Instead, this classifier utilises an ensemble learning method which can be used for classification or also used within a regression. (Indulkar & Patil, 2021) The overarching principle is that the classifier uses multiple ‘trees’ to create a ‘forest’ from which a more accurate prediction can be made.

Such a method of classification harks back to the idea of the, “wisdom of the crowd” first proposed by Aristotle in the 4th Century BC (Landemore,2012). However, the most salient example was demonstrated by Sir Francis Galton. Galton demonstrated that the median of all the 800 guesses to the weight of an ox turned out to be accurate to within 1%. Random Forest Classifier achieves its accuracy by randomly generating decision trees and then applying the same principle demonstrated by Galton.

The trees used in this classifier are based on the binary recursive partitioning trees proposed by Breiman and his colleagues in 1984. Each tree is based upon a random subset of the data and is repeated recursively or until a specific criterion is met. These trees are useful in that they can model complex interactions. Furthermore, they are valuable in that they can scale well for large sample sizes and are robust to outliers that may crop up within the data. (Cutler et al. 2011)

It is here that the ensemble learning method is utilised. Rather than using one tree to make a prediction, the classifier computes numerous trees and uses the average outcome from all the trees to produce its output. In practice this is easiest visualised in Fig.1. In the diagram we can see that the dataset is split into trees, wherein each tree produces their own outcome. These outcomes are then averaged to produce a more accurate result than any one tree individually. Furthermore, with more and more trees, the more accurate the result will be. (Harjadinata & Sibaroni, 2022)

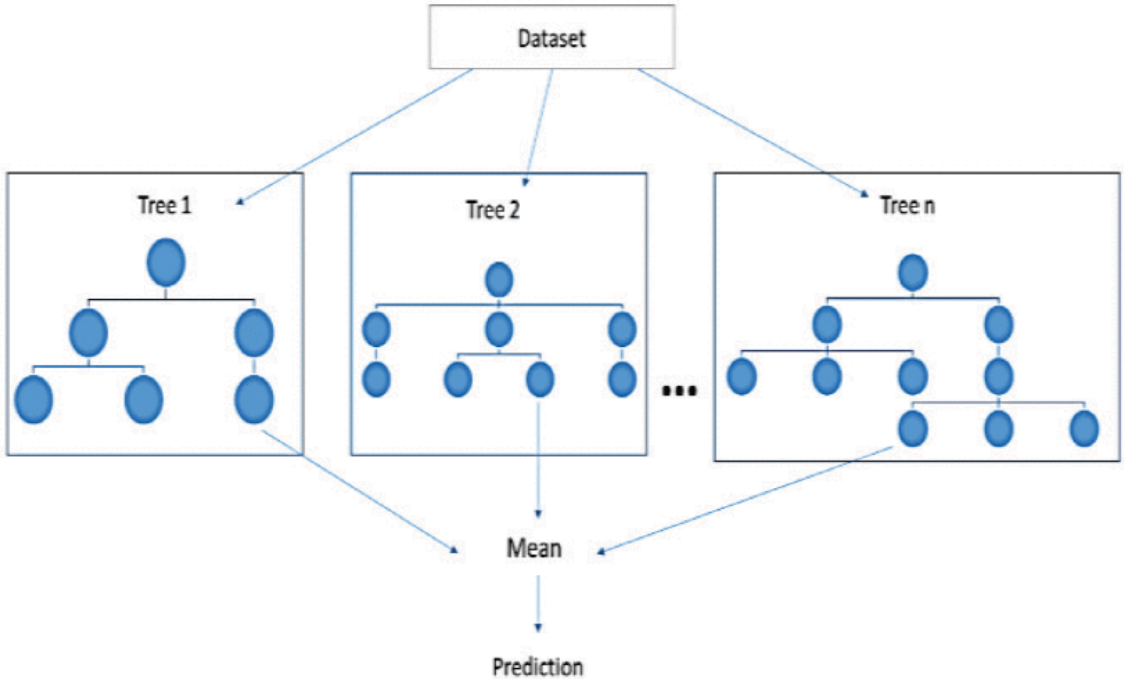

Figure 1. Decision Tree Representation from Root. From Indulkar & Patil, 2021

Random Forest Classifier has some significant advantages over other forms of data classification. One such advantage is its capacity to handle very large datasets and that this process can be relatively easy to train and predict. This capacity is further enhanced by the classifiers ability to handle missing data. This is achieved by the classifiers' use of proximities which can be used to "impute missing values" (Cutler et al., 2011, p18).

3.6 Linear Support Vector Classifier

This classifier is a branch of the Support Vector Machine model. Much like the Naïve Bayes and Random Forest Classifiers, the Linear Support vector Classifier is also a supervised classification model. However, it differs from the other two supervised models in that it relies upon a statistical classification approach, rather than a probabilistic or decision tree approach. The principal idea within Support Vector Classifier is that the model creates a decision plane or 'hyper-plane' that separates the data into two classes (Al Amrani, 2018).

To create this hyper-plane the Support Vector Classifier first isolates the closest data points to the projected hyperplane. These points are referred to as support vectors and are instrumental in determining the best hyper-plane for the model to draw. The distance between each support vector and the projected line is referred to as the 'margin'. The model uses these margins

to find the hyper-plane that has the greatest margin between all the available support vectors. As a result, this creates a hyper-plane that has the greatest separation between the two classes.

To address the fact that the Linear Support Vector generates binary classes a slightly different approach is required. The current project seeks to classify tweets in terms of their sentiment as product of three classes. The model then, will still generate hyper-planes that produce binary classes. However, it will address these in a one-against-one approach, thereby generating three hyper-planes to separate the three classes, namely: positive, negative, and neutral. (Guenther & Schonlau, 2016)

3.7 Model Evaluation

Recall, precision, F-score, and accuracy are common evaluation metrics used in machine learning for classification tasks. They help to measure the effectiveness of a classifier on a specific problem.

1. **Recall**, also known as true positive rate (TPR), recall measures the proportion of actual positive cases that were correctly identified by the classifier. It is given by:

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

   where TP is the number of true positive cases and FN is the number of false negative cases.

2. **Precision** measures the proportion of positive predictions that are actually true positives. It focuses on the quality of the model's predictions and is useful when we want to minimize false positives. It is given by:

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

   where TP is the number of true positive cases and FP is the number of false positive cases.

3. **F-score** is a weighted average of recall and precision, commonly used when there is an uneven class distribution. It is calculated as follows:

$$Fscore = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

The F-score ranges between 0 and 1, with higher values indicating better performance.

4. **Accuracy** this score measures the overall correctness of a classifier's predictions. It indicates the proportion of all instances that were correctly classified by the model. The way to calculate it is the total number of correct predictions divided by the total number of predictions.

$$accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

where TN is the number of true negative instances.

When dealing with multi-class problems, there are several methods in calculating the overall score for each metric.

5. **Micro-averaging** calculates the scores by summing up the individual true positives, false positives, and false negatives from all classes before calculating the metrics. It is given by:

$$average_{mirco} = \frac{TP_{total}}{TP_{total} + FP_{total}}$$

whereIt treats all instances equally and gives higher weight to classes with more samples. Micro-averaged metrics are useful when you care about overall performance and have imbalanced class distribution.

6. **Macro-average** calculates the metric independently for each class and then takes the average. This method treats each class equally and is useful when you want to know how well the model performs on each class. Macro-averaged metrics are suitable when all classes are considered equally important.
7. **Weighted average** is similar to macro-averaging, but it takes into account the proportion of samples in each class. The weighted score is calculated as the weighted average of the per-class performance, where the weight is the number of samples in each class. This method can be

used when you want to prioritize certain classes over others based on their importance or prevalence in the dataset.

To evaluate the best model, this project will use the classification report from sklearn.metrics in python. This is a popular evaluation metric, widely used for model performance in machine learning. It generates a textual report of the classification metrics of a classification model, by taking as inputs the true target labels and predicted target labels. After that it returns precision, recall, F1-score and support related to each label.

The True Positive among all positive predictions are represented by the Precision metrics, whereas recall represents the measure of which the models identify True Positives in all actual positive instances. Finally, the Support column represents the number of instances in each class and the F1-score shows the mean of precision and recall and it gives an overall performance of the model.

3.8 Model Selection and Tuning

Model selection is the process of choosing the best model for a given machine learning task. The aim here is to find the model that best fits the data and also generalizes well to new, unseen data. There are several techniques for model selection, including cross-validation, grid search, and Bayesian optimization. In this project, all supervised models were subjected to K-fold validation in order to calculate their accuracy and grid search was used to optimize the hyperparameters for each model.

Cross-validation involves splitting the available data into training and validation sets, and testing each candidate model on the validation set to see which one performs the best. Information presented in the results section was obtained using 5-fold and 3-fold cross validations.

Hyperparameters are parameters that cannot be learned from data and are set before the training of the model begins, such as the learning rate, regularization strength, number of hidden layers in a neural network, or the choice of kernel function in a support vector machine. Hyperparameter tuning refers to the process of selecting the optimal values for the hyperparameters of a machine learning model.

The goal of hyperparameter tuning is to find the combination of hyperparameter values that results in the best performance of the model on a given task or dataset. This is typically done by evaluating the model's performance using a validation set or through cross-validation, then adjusting the values of the hyperparameters and repeating the process until the best performance is achieved. The process of hyperparameter tuning can help to improve the accuracy and generalization capability of a machine learning model. For the purposes of this project the GridSearchCV class of ScikitLearn library was utilised.

3.9 Ensemble Model and Stacking

Stacking ensemble is a model ensemble technique that involves combining multiple classification or regression models using a meta-classifier or a meta-regressor. The outputs of the base models are used as features for the meta-model to make predictions on the target variable. By combining these models, one can create a more robust and accurate model that is less prone to overfitting than any individual model.

As of normal model training, the stacking process involved splitting the training data into two or more parts. One part was used to train the base models independently, while the other part is used to generate new features or labels by applying each base model on the validation set. These new features were then used to train the meta-model. Logistic regression was used as a meta-model in a stacked ensemble model. Logistic regression is a good choice for a meta-model because it is interpretable, easy to implement, and works well with both categorical and continuous variables. Once the meta-model was trained, it was used to make predictions on the test set. The final prediction is obtained by averaging the predictions generated by all the base models and the meta-model.

Chapter 4: **Results**

4.1 The Dataset

The original data file contained 43943 tweets. A sample of the original dataset is provided in Table 1.

Tweeted	Message	Sentiment
792927353886371840	@tiniebeany climate change is an interesting hustle as it was global warming but the planet stopped warming for 15 yes while the suv boom	-1
793124402388832256	Fabulous! Leonardo #DiCaprio's film on #climate change is brilliant!!! Do watch. https://t.co/7rV6BrmxjW via @youtube	1
793138073542549504	@ShellenbergerMD @DrSimEvans @bradplumer @JigarShahDC should we care about the economics when fighting climate change?	0

Table 1 - Sample tweet sentiment data file

After performing the pre-processing stage, 4730 (11.1%) of the tweets were tagged as duplicates and where consequently removed from the dataset, leaving 39213 tweets for processing.

Table 2 shows a sample of the pre-processing results after removing hashtags, URLs and mentions by replacing them with “URL”, capital letters and converting them to lower case letters.

Tweet ID	Message	Sentiment
792927353886371840	URL climate change is an interesting hustle as it was global warming but the planet stopped warming for 15 yes while the suv boom	-1
793124402388832256	fabulous! leonardo URL film on URL change is brilliant!!! do watch. URL via URL	1
793138073542549504	URL URL URL URL should we care about the economics when fighting climate change?	0

Table 2 Twitter Dataset after pre-processing

Table 3 shows the data after applying tokenization, lemmatization and the removal of stop words.

Tweet ID	Message	Sentiment
792927353886371840	[climate, change, interesting, hustle, global, warming, planet, stopped, warming, 15, yes, suv, boom]	-1
793124402388832256	[fabulous, leonardo, film, change, brilliant, watch, via]	1
793124402388832256	[care, economics, fighting, climate, change]	0

Table 3 - Tokenized Tweets after pre-processing

4.2 Lexicographic Approach

After implementing the steps described in the methodology section, the results are as follows:

Label	Precision	Recall	F1-score	Support
Negative	0.13	0.79	0.23	3789
Neutral	0.21	0.08	0.11	7338
Positive	0.66	0.19	0.29	19785
Macro Average	0.33	0.35	0.21	30912
Weighted Average	0.49	0.24	0.24	30912
Accuracy score: 24%, **Support**: 30912				

Table 4 - Lexicographic Approach Results

As we can see from Table 4, the overall accuracy for the lexicographic approach is only 24 %. Based on the models' own predictions, the best class represented was Positive with 66%. In terms of the how accurate the model was against the pre-identified date it was the Negative class that performed the best with 79%.

4.3 Naïve Bayes Classifier

Across all three classes the model demonstrated a precision between 63-75%. While the recall for all three classes was similar for Negative and Neutral, however the Positive class was the most accurately classified with 94%. The overall accuracy for the model is shown in Table 5 and rests at 73%.

Label	**Precision**	**Recall**	**F1-score**	**Support**
Negative	0.70	0.36	0.47	768
Neutral	0.63	0.39	0.48	1484
Positive	0.75	0.94	0.83	3931
Macro Average	0.70	0.56	0.60	6183
Weighted Average	0.72	0.73	0.71	6183
Accuracy score: 73%, **Support:** 6183				

Table 5 - Naive Bayes Classifier Results

4.4 Random Forest Classifier

For the Random Forest Classifier, the overall accuracy stands at 68 %. Precision for the Negative and Neutral class remain similar, with the positive class demonstrating a precision of 77%. Recall follows a similar pattern with Negative and Neutral again being similar, while the positive class demonstrated an accuracy of 81% as seen in Table 6.

Label	**Precision**	**Recall**	**F1-score**	**Support**
Negative	0.51	0.49	0.50	768
Neutral	0.50	0.44	0.47	1484
Positive	0.77	0.81	0.79	3931
Macro Average	0.59	0.58	0.59	6183
Weighted Average	0.67	0.68	0.67	6183
Accuracy score: 68%, **Support:** 4946				

Table 6 - Random Forest Classifier Results

4.5 Linear Support Vector Classifier (SVC)

In this classifier we can see an overall accuracy of 75%, as seen in Table 7. From the table we can see that the Negative and Neutral classes performed similarly in terms of their precision and recall scores. The Positive class performed the best with a precision score of 81% and a recall of 88%.

Label	**Precision**	**Recall**	**F1-score**	**Support**
Negative	0.60	0.52	0.56	768
Neutral	0.62	0.51	0.56	1484
Positive	0.81	0.88	0.85	3931
Macro Average	0.68	0.64	0.66	6183
Weighted Average	0.74	0.75	0.74	6183
Accuracy score: 75%, **Support:** 6183				

Table 7 - Linear SVC Classifier Results

4.6 Sarcasm and Irony Model Results

4.6.1 Naïve Bayes Results

From Table 8 we can see that this classifier demonstrated a Precision between 58-71%. Similarly, the recall scores also range from 60-64%. Finally, we can see from the table that the classifier demonstrated an overall accuracy of 62%.

Label	**Precision**	**Recall**	**F1-score**	**Support**
irony	0.58	0.62	0.60	4588
regular	0.71	0.53	0.61	3705
sarcasm	0.60	0.69	0.64	4532
Macro Average	0.63	0.61	0.62	12825
Weighted Average	0.63	0.62	0.62	12825
Overall Accuracy: 62%, **Support**: 12825				

Table 8 – Naïve Bayes Classifier Results (Irony/Sarcasm)

4.6.2 Linear Support Vector Classifier Result

For this classifier we can see from Table 9 that the precision of the model across all three classes were similar, in a range between 61-67%. Meanwhile the recall also shows similarity, with a range of 61-68%. The overall accuracy for this model is 63%.

Label	**Precision**	**Recall**	**F1-score**	**Support**
irony	0.61	0.61	0.61	4588
regular	0.67	0.61	0.64	3705
sarcasm	0.63	0.68	0.65	4532
Macro Average	0.64	0.63	0.63	12825
Weighted Average	0.64	0.63	0.63	12825
Overall Accuracy: 63%, **Support**: 12825				

Table 9 – Linear SVC Classifier Results (Irony/Sarcasm)

4.6.3 Random Forest Classifier Results

The Random Forest classifier showed precision scores between 53-63%. For recall the classifier demonstrated and accuracy between 48-71%. As a result, the overall accuracy for this model lies at 58%, as seen in Table 10.

Label	**Precision**	**Recall**	**F1-score**	**Support**
irony	0.61	0.48	0.53	4588
regular	0.53	0.71	0.60	3705
sarcasm	0.63	0.59	0.61	4532
Macro Average	0.59	0.59	0.58	12825
Weighted Average	0.63	0.62	0.62	12825
Overall accuracy: 58%, **Support**: 10260				

Table 10 – Random Forest Classifier Results (Irony/Sarcasm)

4.7 Ensemble Model Results

4.7.1 Climate Change Dataset

Precision score across the classes lie in the range between 63% and 78% according to Table 11. The recall scores for Negative and Neutral are both similar. However, for the Positive class it performed substantially better with an accuracy of 92% The overall accuracy for this Ensemble Model is 75%.

Label	**Precision**	**Recall**	**F1-score**	**Support**
Negative	0.69	0.48	0.56	781
Neutral	0.63	0.45	0.52	1456
Positive	0.78	0.92	0.85	3946
Macro Average	0.70	0.61	0.64	6183
Weighted Average	0.74	0.75	0.73	6183
Accuracy score: 75%, **Support:** 6183				

Table 11 – Ensemble Model (Climate Change)

4.7.2 Sarcasm/Irony Dataset

The data for the Sarcasm and irony dataset are displayed in table 12. Across all three classes we see a precision between 61-68%. Meanwhile the scores for recall lie between 61-67%. The overall accuracy for this Ensemble model then is 64%/

Label	**Precision**	**Recall**	**F1-score**	**Support**
irony	0.61	0.64	0.62	4509
regular	0.68	0.61	0.64	3672
sarcasm	0.66	0.67	0.66	4644
Macro Average	0.65	0.64	0.64	12825
Weighted Average	0.64	0.64	0.64	12825
Overall Accuracy: 64%, **Support**: 12825				

Table 11 – Ensemble Model (Sarcams/Irony)

Chapter 5: **Discussion**

The aim of this paper was to develop a classification model for analysing sentiments towards climate change using Twitter content and test the performance of the approaches in addressing irony and sarcasm.

First, the data from each dataset is prepared according to the steps in the methodology before being passed on to each classifier. For Nave Bayes, Linear SVC, and Random Forest, the data is split as follows: 80% training data and 20% testing. Then the models are trained using a pipeline which includes: text vectorization by using CountVectorizer, text normalization by using TfidfTransformer and model classification. Each model is then evaluated using Precision, Recall, and F1-score. These metrics are used to determine how each model performs overall. At the end of each dataset's performance, without re-training, the ensemble method is run to provide insight into whether the accuracy level could be improved.

Based on the results described earlier, we can see that the Lexicon Approach performed poorly, and this was true for all classes within the Climate Change dataset. The accuracy score of 24% shows that the method only correctly classified 24% of the dataset instances. The Precision score for the Negative class comes out relatively low at 13%. The Precision score for the Positive class is higher at 66%, meaning that the method is better at identifying instances of that class. The Recall score across the classes is between 0.08 and 79%. This indicates that for some reason the model can somewhat correctly classify Negative statements but performs tremendously badly at classifying the other two classes. Such disparity may be due to the fact there is a disparity in the number of tweets present in each class. The Negative tweets numbered far fewer than the other two classes. Finally, the weighted F1-score of 24% means the model performed poorly across all classes. Considering the low performance, this implies that the method has a high false-positive rate and is therefore unable to correctly identify instances of the classes.

The Multinomial Bayes showed that it correctly classified 73% of the instances in the dataset. The Precision and recall score (75% and 94%) belonging to the Positive class is much higher than the Neutral and Negative class. This means that the Naïve Bayes performs better at correctly

identifying instances of the Positive class. Overall, the classifier performed relatively well on the Climate Change dataset, however it needs improvement on the Negative and Neutral class. The reason for that could be the class imbalance. The Linear SVC and Random Forest classifiers were also trained on the same dataset.

The results for the Linear SVC show that this model performed the best among all the classifiers on the climate change dataset. The Random Forest's accuracy score is 68%, which is lower than the Nave Bayes and the linear SVC classifier, whose accuracy scores are 73 % and 75% respectively.

Furthermore, Random Forests’ precision and recall scores are also lower compared to Nave Bayes and Linear SVC. However, its F1-score of 0.79 for the positive class shows that the classifier is able to correctly classify 79% of positive sentiment correctly. Conversely, the negative and neutral classes are much lower (51% and 44%). Overall, the performance of the Random Forest model was not as good as Linear SVC and Naive Bayes, as shown by the metric scores across the classes. This is particularly true for the negative and neutral classes.

In contrast, linear SVC precision, recall, and F1-score scores were generally high for all classes. Precision and recall showed the highest scores obtained for the positive class. This suggests that the classifier is able to make correct predictions for this class while maintaining high precision and recall for the remaining classes. It is worth noting that the Naïve Bayes classifier did perform to almost the same level of accuracy as the Linear SVC classifier on the Climate Change dataset.

In the sarcasm/irony classification task, the lexicographic method was removed based on its performance on the Climate Change dataset. Given how poorly it performed it was deemed unsuitable.

From the performance of the sarcasm dataset, the Linear SVC performed the best with an accuracy of 63%. This accuracy is only a small difference when compared to Naïve Bayes (62%) and Random Forest (58%). The Precision and Recall scores for the Linear SVC report between 61% and 67% across all classes, which is higher than both the Naïve Bayes and Random Forest classifiers. The results show that the model performs relatively well on the irony and sarcasm classes, with F1-scores of 61% and 65%,

respectively. The Recall and F1-score of 61% and 64% shows that the Linear model is better at identifying regular tweets. Whereas the Naïve Bayes Recall of 69% and F1-score of 64% shows that it identifies the sarcasm class better than both Regular and Irony classes. Random forest scored highly when identifying the Regular class based on its Recall score (71%) but performed worse on both other classes.

Comparing the two datasets it is noted that the performance scores across the classes appear to be more balanced for the Sarcasm/Irony dataset. Whereas in the Climate Change dataset it is noted that all classifiers performed better at classifying the Positive sentiment. When examining Figure 2. containing the distribution of classes, we can see that the Positive class contains significantly more tweets than the Neutral and the Negative class. AS there are so many more tweets in the positive class it many artificially inflate the results.

After evaluating each dataset this project uses an ensemble method which is explained further in the Methodology section. This is a method of combining two classifiers with the aim to produce even more accurate results. The two best performing classifiers are chosen for this task, namely, Linear SVC and Naïve Bayes.

With respect to the Climate Change dataset, we can see that the ensemble method did not perform better than any one individual classifier. The Linear SVC classifier produced an accuracy of 75% which is the same as the accuracy produced by the ensemble method. However, it scored higher on the Recall score when identifying the Positive class (92%). This may reflect the fact that the Naïve Bayes Classifier produced a Recall score for the Positive class of 94%. The ensemble methods Recall for the Negative and Neutral class performs worse than the best performing classifier alone, namely Linear SVC. However, this may be due to the fact that the Naïve Bayes classifier is reducing the accuracy in the ensemble as it performed poorly when classifying Negative and Neutral classes. It is again worth noting that the fact that all three supervised classifiers scored much higher when classifying Positive sentiment may be partly based on the larger number of positive sentiments in the dataset.

The ensemble method for the Sarcasm/Irony dataset however, scored only 1% higher when compared to the best performing classifier. The ensemble

produced an accuracy of 64% compared to the Linear SVC classifier which produced an accuracy of 63%. That being said, we can see that the Linear SVC seemingly performs better in isolation than when incorporated into the ensemble. The Naïve Bayes performed somewhat worse than the Linear SVC and may be pulling Recall scores down. Despite this however, the improvement seen suggests that by perhaps including other classifiers into the ensemble it may produce even greater accuracy.

The ensemble approach that was undertaken in this study is similar to the procedure performed by Al Ahmrani and their colleagues (2018). In their study they combined Random Forest and Support Vector Machine to produce greater accuracy than each of the classifiers individually. The Random Forest scored an accuracy of 81 %, the Support Vector Machine scored 82.4 % and the ensemble produced an accuracy of 83.4%. The current study then, also demonstrates that the combination of two classifiers does in fact produce a more accurate result, at least with reference to the Sarcasm/Irony dataset. Furthermore, Godara and her colleagues (2021) utilising an even more significant ensemble approach for detecting sarcasm. The authors used a few ensemble methods combining machine learning algorithms such as KNN, Naïve Bayes and Support Vector Machine. In their results they reported very high accuracy as high as 99.17% in overall accuracy. Further demonstrating that ensemble methods may represent a better path to addressing the problem of accuracy in Sentiment Analysis.

.

Chapter 6: **Conclusions**

This research has been concerned with analysing and comparing the performance of one unsupervised and three supervised machine learning algorithms. After the evaluation of each classifier, an ensemble method for each dataset was used. This ensemble combined the two best performing algorithms, Naïve Bayes and Linear SVC.

This work demonstrates that machine learning algorithms in combination with ensemble methods can produce accurate results when predicting sentiment. According to a study Pew research Center, 72% of adults use social media of some type, as of 2021. As such the capacity to glean information about the sentiment of large volumes of social media data is invaluable to institutions and businesses. By refining the accuracy of Sentiment Analysis, it can help institutions and businesses have a deeper understanding of their users' opinions. For example, with Climate Change, a more accurate Sentiment Analysis of social media data may help institutions to address how their message is reaching the wider public. Similarly, figurative language is often expressed on social media. Therefore, having a means to classify sarcastic or ironic social media text with greater accuracy ensures that the nuanced messages continued within figurative language are not lost in the process.

Based on the results of this research and that of Godara and her colleagues (2021), it can be proposed that ensemble methods seem to produce the best results with respect to accuracy. However, surely the more complicated an ensemble method becomes the greater the limitations that may occur. Computational capacity for example may become an issue when many classifiers are combined. Similarly, there can be a problem of overfitting in such methods. In this research it became apparent that there was overfitting happening with the Random Forest classifier before its hyperparameters were tuned, as it reported 100% accuracy. It is therefore likely that combining such a method into an ensemble may produce some overfitting. Finally, it is worth noting that ensemble methods are not the simplest, which may impact their implementation and interpretability.

Future work to improve accuracy may include the use of even larger datasets. Making use of larger datasets means that the models have more

experience of language and how it is used. This may be of particular importance when identifying figurative language. Furthermore, the use of larger datasets may help to negate some of the impact of overfitting as a smaller dataset may lead to the model generalising poorly to new data. Another aspect to consider is the fact that the Climate Change dataset used in this study was not particularly balanced, so using datasets with more even class distributions may produce better accuracy. Other more advanced technology can also improve the accuracy like Recurrent Neural Network which was demonstrated in a recent study to be highly accurate (Topbaş, 2021).

Chapter 7: **References**

Al Amrani, Y., Lazaar, M. and El Kadiri, K.E. (2018). Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Procedia Computer Science, 127, pp.511–520. doi:https://doi.org/10.1016/j.procs.2018.01.150.

Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth, New York (1984).)

Cutler, A., Cutler, D.R. and Stevens, J.R. (2012). Random Forests. Ensemble Machine Learning, [online] pp.157–175. doi:https://doi.org/10.1007/978-1-4419-9326-7_5.

eMarketer, Insider Intelligence. (2022). Number of Twitter users worldwide from 2019 to 2024 (in millions). Statista. Statista Inc.. Accessed: February 27, 2023. https://www.statista.com/statistics/303681/twitter-users-worldwide/

Fernández Gavilanes, Milagros & Àlvarez-López, Tamara & Juncal-Martínez, Jonathan & Costa-Montenegro, Enrique & González-Castaño, Francisco. (2015). GTI: An Unsupervised Approach for Sentiment Analysis in Twitter. 533-538. 10.18653/v1/S15-2089.

Gamallo P., Garcia M., Santiago, “Tass: A naivebayes strategy for sentiment analysis on spanish tweets,” in International Conference on Social Informatics, pp. 215-221, 2012.

Godara, Jyoti & Aron, Rajni & Shabaz, Dr. Mohammad. (2021). Sentiment analysis and sarcasm detection from social network to train health-care professionals. World Journal of Engineering. ahead-of-print. 10.1108/WJE-02-2021-0108.

Godara, J., Batra, I., Aron, R. and Shabaz, M. (2021). Ensemble Classification Approach for Sarcasm Detection. Behavioural Neurology, 2021, pp.1–13. doi:https://doi.org/10.1155/2021/9731519.

Greenpeace UK (2022) What is the UK doing about climate change? Available at: https://www.greenpeace.org.uk/challenges/climate-change/what-is-the-uk-doing-about-climate-change/.

Guenther, N. and Schonlau, M. (2016). Support Vector Machines. The Stata Journal: Promoting communications on statistics and Stata, 16(4), pp.917–937. doi:https://doi.org/10.1177/1536867x1601600407.

Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International AAAI Conference on Weblogs and Social Media

Indulkar, Y. and Patil, A. (2021). Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis. [online] IEEE Xplore. doi:https://doi.org/10.1109/ESCI50559.2021.9396925.

Jackson, R. (2022) The Effects of Climate Change. Available at: https://climate.nasa.gov/effects/.

Kelleher, J.D. and Tierney, B. (2018). Data science: Introduction. Cambridge, Massachusetts ; London, England: The MIT Press.

Klinger, R. (n.d.). *Irony Sarcasm Analysis Corpus*. [online] Roman Klinger’s Homepage. Available at: https://www.romanklinger.de/ironysarcasm/ [Accessed 9 May 2023].

Landemore, Hélène (2012). "Collective Wisdom—Old and New" (PDF). In Landemore, Hélène; Elster, Jon *(eds.). Collective wisdom: principles and mechanisms.* Cambridge, England*:* Cambridge University.

Leung, K. (2007). Naive Bayesian Classifier. [online] Available at: https://cse.engineering.nyu.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf.

Liu, B. (2020) “Introduction,” in Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. 2nd edn. Cambridge: Cambridge University Press (Studies in Natural Language Processing), pp. 1–17. doi: 10.1017/9781108639286.002.

Mishne G., “Experiments with Mood Classification in Blog Posts”, Proceedings of 1st Workshop on Stylistic Analysis of Text for Information Access, 2005

Mohammad, S.M. (2017). Challenges in Sentiment Analysis. A Practical Guide to Sentiment Analysis, pp.61–83. doi:https://doi.org/10.1007/978-3-319-55394-8_4.

Neviarouskaya A., Prendinger H. and Ishizuka M., "SentiFul: Generating a reliable lexicon for sentiment analysis," 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, Netherlands, 2009, pp. 1-6, doi: 10.1109/ACII.2009.5349575.

Nielsen, F.Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. [online] arXiv.org. Available at: https://arxiv.org/abs/1103.2903.

Office for National Statistics (ONS), released 28 October 2022, ONS website, article, Worries about climate change, Great Britain: September to October 2022

Oxford Dictionary (2022). Oxford learner’s dictionaries. [online] Oxfordlearnersdictionaries.com. Available at: https://www.oxfordlearnersdictionaries.com/.

Pang, B., Lee, L. and Shivakumar Vaithyanathan (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. arXiv (Cornell University).

Peter and Turney (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Meeting of the Association for Computational Linguistics, pp.417–424.

Pozzi, F.A. et al. (2016) Sentiment Analysis in Social Networks. Morgan Kaufmann.

Pew Research Center (2021). Social Media Fact Sheet. [online] Pew Research Center. Available at: https://www.pewresearch.org/internet/fact-sheet/social-media/.

Saifullah, S., Fauziyah, Y. and Aribowo, A.S. (2021). Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. Jurnal Informatika, 15(1), p.45. doi:https://doi.org/10.26555/jifo.v15i1.a20111.

Shajahan, S. and Poovizhi, T. (2022). A Novel Approach to Estimation Precision and Recall for Star Rating Online Customers Based on Negative Hotel Reviews using Multinomial Naive Bayes over Multischeme Classifier.

Singh, V. et al. (2018) “Sentiment analysis using lexicon based approach,” 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) [Preprint]. Available at: https://doi.org/10.1109/pdgc.2018.8745971.

Stanford University (2008), Stemming and lemmatization. Available at: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, [online] 37(2), pp.267–307. doi:https://doi.org/10.1162/coli_a_00049

Topbaş A., Jamil A., Hameed A. A., Ali S. M., Bazai S. and Shah S. A., "Sentiment Analysis for COVID-19 Tweets Using Recurrent Neural Network (RNN) and Bidirectional Encoder Representations (BERT) Models," 2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, Pakistan, 2021, pp. 1-6, doi: 10.1109/ICECube53880.2021.9628315.

Turney P.D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; Association for Computational Linguistics; 2002:417–424.

www.kaggle.com. (n.d.). *Twitter Climate Change Sentiment Dataset*. [online] Available at: https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset.

Yang C., Lin K. H. Y. and Chen H. H., "Emotion Classification Using Web Blog Corpora," IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), Fremont, CA, USA, 2007, pp. 275-278, doi: 10.1109/WI.2007.51.

Zahoor S. and Rohilla R., "Twitter Sentiment Analysis Using Machine Learning Algorithms: A Case Study," *2020 International Conference on Advances in Computing, Communication & Materials (ICACCM)*, Dehradun, India, 2020, pp. 194-199, doi: 10.1109/ICACCM50413.2020.9213011.