Module Code & Module Title

CU6051NT - Artificial Intelligence

Assessment Weightage & Type

25% Individual Coursework

Year and Semester

2021-22 Autumn

Student Name: Prashant Nepal

London Met ID: 19033562

College ID: NP05CP4S200019

Assignment Due Date: 22th December

Assignment Submission Date: 22th December

Academic Supervisor: Mr. Prateek Kokh Shrestha

Word Count: 6385

I confirm that I understand my coursework needs to be submitted online via Google Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission, and a mark of zero will be awarded.

Table of Contents

1. Introduction *1*

1.1. Explanation of the AI topic/concepts used. *1*

1.2. Explanation of Chosen Problem Domain/Topic *3*

2. Background *5*

2.1. Research Work on Chosen Topic/Problem Domain *5*

2.2. Review and Analysis of the Existing Works *8*

3. Solution *11*

3.1. Explanation of the Proposed Solution *11*

3.2. Explanation of the Algorithm Used *14*

3.3. Pseudocode for the solution *20*

3.4. Flowchart of the solution *22*

4. Conclusion *23*

4.1. Analysis of the Work Done *23*

4.2. How the solution addresses the real-world problems *25*

4.3. Further work *26*

References and Bibliography *27*

Table of Figures

Figure 1: Naïve Bayes Classifier for Sentiment Analysis *13*

Figure 2: Flow chart *22*

Table of Tables

Table 1: Labeled training data 16

Table 2: Frequency of words A.. 16

Table 3: Frequency of words B.. 16

1. Introduction

1.1. Explanation of the AI topic/concepts used.

Homo Sapiens, an intelligent organism, are competent enough to make rational decisions. For an extended period, efforts have been made to realize how a human being thinks; like, how a handful of matter can predict, perceive, understand, and manipulate a universe, far deeper and more intricate than itself. Artificial Intelligence not only attempts to acknowledge it but moves miles for developing intelligent objects.

Alan Turing's publication, "Computing Machinery and Intelligence," in 1950, evolved the concept of AI, where he suggested a deterministic path for distinguishing machine intelligence from that of the human (Turing Test) (Mueller & Massaron, 2021). However, the term AI was coined in 1956 at Dartmouth College during a conference. During the conference, scientists like Marvin Minsky had presented their opinion about AI's bright future. But in the mid-1970s, due to sluggish progress and several reports, the government dropped funding for AI research. Then again, the British government revived AI research by initiating financing in the 1980s. Later in 1997, when IBM's Deep Blue defeated world chess champion, a Russian grandmaster, Garry Kasparov, AI became the matter of conversation in every corner of the world and quickly caught the track of rapid development (Lewis, 2014).

An imparting cognitive ability of a machine can be determined as artificial Intelligence, whose benchmark has been compared to human Intelligence regarding speech, vision, and reasoning (Rungta, 2018). As a result, AI is divided into three levels:

- Narrow AI – where a machine can perform a few tasks better than a human.
- General AI – where a system can perform any task as accurately as a human.
- Active AI – where a machine can perform and beat humans in many jobs.

Machine learning (ML), a subset of AI, depends upon the underlying algorithms for analyzing enormous datasets. ML trains computers for automating its task and makes conclusions with minimal human interventions by identifying specific data patterns. ML provides data to the algorithm for the learning mechanism, predicting the output of new input once the training is completed.

Lastly, deep learning, a subset of machine learning, use a deep neural network (like in the human brain) to create deeper layers for learning from the data. The deeper (hidden) layers are stacked in the neural network architecture.

After getting acquainted with the fundamental concepts of AI, the essence of implicating it in the real-life can be speculated. In the current scenario, humans have embraced AI in different sectors like astronomy, healthcare, finance, social media, data security, robotics, agriculture, education, etc.

Ultimately, AI has diverse fields of study that include Machine Learning, Deep Learning, Neural Networks, Computer Vision, Robotics, Natural Language Processing, etc. Moving forward in this project, which is based on Natural Language Processing (NLP), the discussion related to the specific topic will be delivered extensively.

Natural Language Processing scrutinizes computer usage for processing and understanding human language for achieving practical tasks. NLP merges the concept of cognitive science, computing science, computational linguistics, and Artificial Intelligence into a single platform. From the researchers' viewpoint, NLP aims to model the cognitive mechanisms through the production and acknowledgment of human languages. On the other hand, from engineers' perspective, NLP is focused on developing systems that assist human language interaction with machines. It is a system specifically designed to convey natural semantics (meaning) through discrete or symbolic systems (Deng & Liu, 2018). The primary sector of NLP includes lexical analysis, information retrieval, knowledge graph, parsing, question answering, speech recognition, dialogue systems, natural language generation, spoken language understanding, sentiment analysis, and more.

1.2. Explanation of Chosen Problem Domain/Topic

Sentiment analysis or opinion mining, a field of NLP, concerns extracting underlying subjective knowledge from the lines of text provided. The research regarding sentiment analysis has been skyrocketed in the last decade, as the competition involving brands to learn social sentiment upon their business, service, and product upsurge (Gupta, 2018). Also, the study for understanding the motive of social media users upon political issues, current affairs, elections, social issues, and online shopping by analyzing their online conversations is gaining immense attention.

The information that is available over the internet is either facts or opinions. Facts are objective, providing information about the entities without any sentiments. On the other hand, opinions are subjective, explaining people's emotions towards any event or entity (Agarwal & Mittal, 2016). The enormous amount of data produced in consumers' arguments, views, opinions, and emotions related to brands, products, events, and politics significantly influence readers, politicians, and product vendors. By managing and analyzing unstructured data from online conversations, reviews, posts, and status, assumptions can predict user sentiments that significantly influence the subject matter.

On 6th January 2021, mass supporters of former President Donald Trump strike the United States Capitol located in Washington D.C. The mob demanded to overturn the election, which they believed was rigged; President Joe Biden won that. Unfortunately, five people lose their lives in that hideous action, including a police officer. The event was followed by the suicidal death of four more officers, being traumatized by the incident (Peterson, 2021). The riot was thoroughly planned using social media platforms, i.e., Facebook, where people from different parts of the country participated (Timberg et al., 2021). Now it can be concluded that the incident was entirely avoidable. The spreading of false news and planning of dreadful events were completely preventable using sentiment analysis.

Not only limiting to stop circulating misinformation, but the sentiment analysis can also be implemented for various purposes like evaluating people's emotions, attitudes, and appraisals toward entities and attributes available in a different form of text. The entities could be issues, topics, services, products, organizations, or individuals (Liu, 2015). Ultimately, the assumptions can be advantageous for increasing the quality of products or services, providing comfort to the customers, generating statements on several topics, spreading business, and more that could uplift the welfare of an organization.

Currently, implications of sentiment analysis have popularized in mitigating various problem domains, i.e., health care, financial services, stock market prediction, political elections, tourism, and hospitality. The result served from these applications help startups, and the government regulates proper and well-defined regulations. Although sentiment analysis has increasingly been conducted, research on handling biased sentences is limited due to substantial complications in this field.

2. Background

2.1. Research Work on Chosen Topic/Problem Domain

Intensive research has been carried out for understanding the implementation of sentiment analysis, its limitations, and its applications. While going through several resources, forums, journals, books, and online mediums, it was found that sentiment analysis can be achieved through different approaches in Machine Learning and AI. For instance, Support Vector Machine (SVM), Naïve Bayes Classifier, RNN (Recurrent Neural Networks), regional CNN-LSTM, GRU (Gated Recurrent Units), linear regression, Maximum Entropy Classifier, Decision Tree Classifier, and so on (Medhat et al., 2014).

Several types of sentiment analysis, like fine-grained sentiment analysis, serve a precise division level by categorizing opinions into 5-star scale, i.e., from very positive to very negative. Emotion detection depicts the stages of anger, shock, sadness, happiness, and frustration rather than negativity and positivity—next, intent-based analysis, which speculates actions and opinions from the written text—for instance, taking specific activity for the frustration shown by customers over online comments. And the last one is an aspect-based analysis that serves the prediction made for specific product components rather than the whole entity (TechTarget, 2021). For this project, fine-grained sentiment analysis will be performed where the opinions will be categorized into three labels, i.e., positive, negative, and neutral.

The workload of sentiment analysis can be achieved through two different methods that are rule-based approach and automatic sentiment analysis. In the rule-based system, the statement corpus will be treated through various processes that involve stemming, tokenization, part of speech tagging, parsing, and lexicon analysis. The whole group of words will be divided into negative and positive groups. The applied algorithm figures out the criteria and classifies the word as positive or negative based on their polarity (Hutto & Gilbert, 2014). Although it serves results but lacks precision and flexibility, they do not tend to be handy.

On the other hand, automatic sentiment analysis encompasses machine learning techniques for analyzing sentiments of provided gist of the information. Automation in sentiment analysis increases precision and accuracy, which lowers the chances of complications. This approach implies supervised machine learning algorithms for classification and may also use unsupervised machine learning algorithms to explore the data (Boiy et al., 2007). The classification algorithms for this approach may include linear regression, SVM, RNN, Naïve Bayes, CNN-LSTM, and so on.

Support Vector Machine (SVM) is one of the most capable and powerful approaches for sentiment analysis, which efficiently detects textual polarity. It is a supervised machine learning algorithm that can be used for classification or regression problems. The classification will be performed by figuring out the hyper-plane that distinguishes classes been plotted in an n-dimensional graph, which is drawn with the help of Kernels (mathematical functions). Collecting appropriate data for training and testing, vectorizing data, and developing linear SVM models for training and predicting are the steps for building excellent models (Reddy, 2018).

CNN (Convolution Neural Networks) – LSTM (Long Short-Term Memory) is another practical approach for fine-grained sentiment analysis using deep learning. CNN utilizes a statement as a region from where the relevant information is extracted and weighted based on their contributions. Now, the regional knowledge will be integrated across areas using LSTM, and by combining them, both the sentence information and textual dependency will be predicted (Wang et al., 2016). Although these algorithms outperform other machine learning approaches, they require substantial computational power, dataset, and time.

Another popular machine learning algorithm used for sentiment analysis is Naïve Bayes Classifier based on Naïve Bayes (NB) theorem, a probabilistic machine learning algorithm. The occurrence result will be predicted by calculating the conditional probability of an event given another event as a condition. Naïve Bayes Classifier is also divided into other types like; Multinomial Naïve Bayes that classifies the problem based upon its category. Bernoulli Naïve Bayes predicts the Boolean variables with only two parameters as 'yes' or 'no.' Finally, Gaussian Naïve Bayes classifies the continuous data rather than discrete ones (Gandhi, 2018).

Since the existence of one feature does not affect another, this theorem is called naïve. As a result, it may hinder the system's performance in real-life scenarios where predictors are dependent on each other. This type of classifier is widely popular for spam filtering, recommendation systems, and sentiment analysis.

In a nutshell, sentiment analyses are tricky problems requiring heavy tasks rather than routine extraction. The definition of context and polarity must be well defined using text vectorization, which maps word connections and their relations. Tools like doc2vec and word2vec could facilitate gaining the most out of the model. Further, for subjectivity and tone determination, the characterization of the product should be considered. The model should be trained with diverse corpus and deep context for handling sarcasm and irony identification. Also, by marking polar messages, the model could define the neutral tone. Eventually, through these processes, a well-refined NLP system could be developed.

2.2. Review and Analysis of the Existing Works

Sentiment analysis determines people's appraisal of several entities in terms of different matrices like positive, negative, neutral, etc. The act of realizing the importance of sentiment analysis and keeping their records can be traced back to the mid-twentieth century. Along with the digitalization, the paper works have shifted into computer software, and analysis of sentiments is done through machine learning processes. Large systems have been developed to deal with sentiment analysis; meanwhile, research is also running to increase their efficiency. A few projects related to this topic are reviewed and critically analyzed in this section.

The paper released by Piek Vossen and Isa Maks, "A Lexicon Model for Deep Sentiment Analysis and Opinion Mining Applications," depicts the thorough processes involved in developing and evaluating complicated lexicon models. The model categorizes words based on adjectives, nouns, and verbs for sentiment analysis jobs. This approach has contributed to advancing research to create resources required for sentiment analysis. At first, the opinions presented in textual form have been contextualized using several holders and opinions' targets utilizing the semantic FrameNet model. Also, the existence of subjective relation in-between different actors are considered. Secondly, semantic categorization has been used to create words' annotation, which are equivalent to opinion mining but dependent on speaker and writer. In addition, the word annotation model facilitates conveying the text corpus's sentiments and emotions. At last, the model also holds the role of the writer and speaker as delivered in the text, which can be crucial for accurate sentiment analysis and detecting bias from the context.

Another work of Josef Steinberger in " provides a solution for the problems faced while obtaining lexical resources from various languages for sentiment analysis. This project contributes to enhancing the state of multilingual sentiment analysis. The proposed method, 'triangulation,' facilitates creating resources for two different languages and translating into other languages. This method aims to dissipate the noise created from automatic translation and the problem of ambiguity in word sense. On the top, the author has mentioned expanding, correcting, and improving the lists of words for implementing an exact sentiment analysis system.

"On Developing Robust Models for Favorability Analysis: Model Choice, Feature Sets and Imbalanced Data" is an article by Peter C.R.Lane, which points out the difficulties in retrieving opinions from media contents, i.e., figuring documents having positive or negative approvals. The experimentation conducted over real-time media data for sentiment analysis has been provided through this project. However, the balance between positive and negative documents' classes and fluctuating content may lead to the underperformance of ML models. By addressing such issues, several alternatives have been proposed that can be implemented based on the characteristics and tasks of the dataset.

Another article by Antal van den Bosch and Matje van de Camp, "The Socialist Network," highlights a sentiment analysis system that extracts personal social networks from ancestral texts. The core contribution of this project is that it addresses the various challenges by identifying them in sentiment analysis. Although the outcome will be enormously dependent upon analyzing opinions, authors have depicted high ML algorithms, provided healthy comparative vision, and usage of different properties and linguistic resources. Hence, the obtained results are valuable for further enhancement in sentiment analysis.

"Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models" is an article by Wei Li et al. demonstrating the CNN-BiLSTM and CNN-LSTM models. The sentiment padding method introduced in the article assists in developing consistent-sized data samples and enhancing the size of sentiment information. Also, the issues related to gradient vanishing between hidden and input layers have been tackled. The proposed loss function assists two branches of a model to be trained at different speeds, which eventually increases the performance of sentiment analysis. The high performance of the CNN-LSTM model over the provided dataset concludes that the system's performance will depend upon the structure of sentences. The skip connection operation decreases NN models' performance and training speed to complete. Further, the factors involved in creating coupling between two branches of the model must be researched.

The overall research conducted and delivered above has carried out sentiment analysis using different complex approaches. Several issues have been encountered while integrating proposed models into the datasets. As a result, the authors also explained the solution for handling such shortcomings. Ultimately, sentiment analysis is a challenging topic, which is being flourished in several fields. Intensive further works must be conducted for blindly believing in these systems. Further, the study of models is limited to detect explicit, direct sentiment expression and implicit expressions, which are contained in sentences through emotion eliciting, objective arguments.

For this project, the Naïve Bayes theorem has been implemented to analyze textual content sentiments. The latter section of the document discusses the benefits, shortcomings, challenges, and space for improvement using the Naïve Bayes algorithm.

3. Solution

3.1. Explanation of the Proposed Solution

Along with the increase in internet usage, clients' interaction with web applications has surged rapidly. The amount of data accumulated from users' reviews, comments, ratings, forms, and posts are evaluated and employed. Now, the analytics figure can bring changes that benefit both organization and clients. Analyzing the sentiments can be performed from different approaches like Machine Learning and the rule-based system. The Machine Learning approach is widely adapted does not require to be limited based on classification rules (Ravi & Ravi, 2015).

In the ML approach, predefined training data will be fed to an algorithm, which then learns to classify the input data into various labels. While going through deep research on sentiment analysis, multiple algorithms were considered for the implementation. Among several algorithms, being straightforward to understand and implement, the Naïve Bayes approach has been adopted to classify sentiments. Through this approach, each extracted feature is treated as independent of others.

The tweets collected from an online source will be fed to the Naïve Bayes Classifier for training and testing purposes. The data provided to the system will be ready to pre-process, train, and test by importing necessary libraries. After importing training data into the model, data pre-processing will be carried out to remove unnecessary punctuation marks, tabs, and emoticons from texts. Next, defining labels and extracting features from the dataset will be performed.

Further, a bag of words will be created that stores unique dishes from the whole training model to make phrases appear in specific statements. During this approach, the frequency of words' appearance will be considered. However, the position of a text in a sentence will be ignored, due to which it lacks in solving complexly structured sentences. Then for training, the dataset column will be considered a statement and label.

After creating a bag of words, words from the list will be converted into the frequencies of their repetition. The actual Naïve Bayes theorem will be implied to calculate the conditional probability. Eventually, the model will independently predict the statements as negative, positive, or neutral by considering the highest possibility.

Also, the tokenization and stemming of data from the dataset could result in higher accuracy. This approach trims down the long text into precise smaller chunks of words. Also, utilizing lemmatization and stop words techniques increases the probability of gaining better accuracy results. As mentioned earlier, tools like doc2vec and word2vec can transform textual information into vector form. These are the standard practices that are prevalent in modern NLP research.

The Naïve Bayes Classifier is based upon the Bayes Theorem that demonstrates the conditional probability for assuming the predicted P(A | B) event, given that event B has occurred. The equational representation of the formula along with the notation is depicted below:

$$\mathrm{P(A \mid B)} = \frac{\mathrm{P(B \mid A)} \times \mathrm{P(A)}}{\mathrm{P(B)}}$$

Where:

A, B = events

P(A) = independent probability of an event A’s occurrence (prior)

P(B) = independent probability of an event B’s occurrence (evidence)

P(A | B) = probability of A such that B is prior condition (posterior)

P(B | A) = probability of B such that A is prior condition (likelihood)

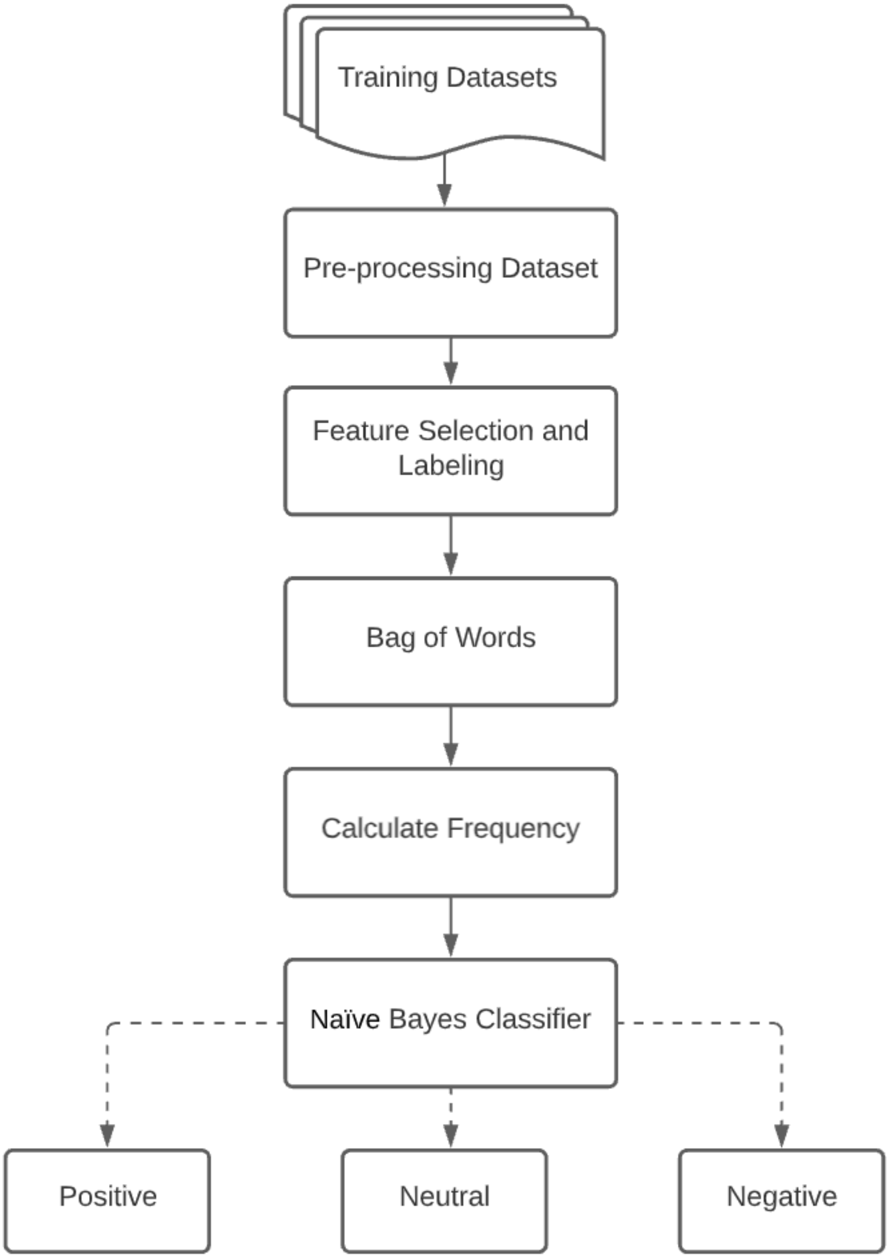

Figure 1: Naïve Bayes Classifier for Sentiment Analysis

3.2. Explanation of the Algorithm Used

Naïve Bayes, a probabilistic ML algorithm, is based on the Bayes Theorem. The rapid development of a robust ML model can be achieved from this approach. As this is a supervised learning algorithm, it requires an appropriate dataset for training. This algorithm calculates the probability from the text frequencies assigned to each value in the training dataset. Finally, after calculating probabilities for each word concerning a given condition, like: if the word is negative, positive, or neutral, the output label will have the highest probability.

After importing necessary libraries, the relevant dataset will be imported where pre-processing techniques will be implied to eliminate unnecessary punctuation, tabs, and emoticons from the text. After extracting features and labels (positive, negative, and neutral), the bag of words will be created to determine the frequency of word repetition. Next, depending upon the respective labels, the independent probabilities for every word will be calculated. Ultimately, the trained model from the provided dataset will prepare relevant output.

Without eliminating the core concept, the Naïve Bayes formula has been tweaked and twisted to increase the model's efficiency. The modified version of the Naïve Bayes formula is portrayed below:

$$P(y \mid x) = \frac{P(x \mid y) \times P(y)}{P(x)} \approx P(x \mid y) \times P(y)$$

Since the conditional probability of other labels will be compared for figuring out the highest probability, the denominator of the equation can be eliminated. In the above equations, ' y' is the features, whereas 'x' is the label, i.e., negative, positive, or neutral. For several features, it can be presented as:

$$P(y \mid x) = P(a_1, a_2, ... a_n \mid y) \times P(y)$$

Where,

$$P(a_1, a_2, ... a_n \mid y) \times P(y) = P(a_1 \mid y) \times P(y) \times P(a_2 \mid y) \times P(y) ....P(a_n \mid y) \times P(y)$$

Now, by calculating the probability of each word for labels, conditional probability can be calculated from this approach. On further modifying the formula,

$$P(W_k \mid \text{label}) = \frac{nk+1}{n+|\text{Vocabulary}|}$$

The above equation tends to calculate the probability of W_k for the given label (negative, positive, or neutral). The n in the given equation represents the number of words presented in the dataset's negative, positive, or neutral conditions. On the other hand, nk represents the frequency of words in each of the labeled states. $|Vocabulary|$ in the above equation is the cumulative number of words in the vocabulary, i.e., a bag of words.

If the number of words in the vocabulary tends to be 0, the above formula could lead us in the wrong direction. So, to prevent ∞ as an outcome of the probability, $|Vocabulary|$ has been added in the denominator of an equation. Next, if the frequency of the word in the given label is 0, the above formula could yield 0 probability. So, to prevent this shortcoming, one has been added to the numerator of the equation.

Finally, the probability for every word based upon its labels will be calculated from the training dataset. After training, the model the system will be ready to predict whether the given input is positive, negative, or neutral based on the most significant probability that each one has.

An appropriate example has been highlighted below to understand the algorithms' workflow properly.

Statements	**Labels**
You are awesome.	Positive
Thank you.	Positive
Will be back.	Neutral
You are so dumb	Negative
Don't like you	Negative
You are jealous.	Negative

Table 1: Labeled training data

Now the bag of words will be:

['you', 'are', 'awesome', 'thank', 'will', 'be', 'back', 'fell', 'asleep', 'don’t’, ‘like’, ‘jealous’]

Then the words concerning their frequency will be:

Statements	You	Are	Awesome	Thank	Will	Be	Back	so	**Labels**
You are awesome.	1	1	1	0	0	0	0	0	Positive
Thank you.	1	0	0	1	0	0	0	0	Positive
Will be back.	0	0	0	0	1	1	1	0	Neutral
You are so dumb	1	1	0	0	0	0	0	1	Negative
Don’t like you	1	0	0	0	0	0	0	0	Negative
You are jealous.	1	1	0	0	0	0	0	0	Negative

Table 2: Frequency of words A

Statements	Dumb	Don’t	Like	Jealous	**Labels**
You are awesome.	0	0	0	0	Positive
Thank you.	0	0	0	0	Positive
Will be back.	0	0	0	0	Neutral
You are so dumb	1	0	0	0	Negative
Don’t like you	0	1	1	0	Negative
You are jealous.	0	0	0	1	Negative

Table 3: Frequency of words B

Now, the probability of positive, negative, and neutral is:

P(positive) = 2/6

P(neutral) = 1/6

P(negative) = 3/6

For calculating the probabilities for entire labels, the formula will be:

$$P(W_k | \text{label}) = \frac{\text{nk+1}}{\text{n+|Vocabulary|}}$$

<u>At first, for **positive** sentiments:</u>

Total number of words in positive sentences (n) = 5

Total number of words in vocabulary ($|Vocabulary|$) = 12

P(you | positive) = $\frac{2+1}{5+12}$ = 0.18

P(are | positive) = $\frac{1+1}{5+12}$ = 0.12

P(awesome | positive) = $\frac{1+1}{5+12}$ = 0.12

P(thank | positive) = $\frac{1+1}{5+12}$ = 0.12

P(be | positive) = $\frac{0+1}{5+12}$ = 0. 06

P(so | positive) = $\frac{0+1}{5+12}$ = 0.06

P(don't | positive) = $\frac{0+1}{5+12}$ = 0.06

P(like | positive) = $\frac{0+1}{5+12}$ = 0.06

P(jealous | positive) = $\frac{0+1}{5+12}$ = 0.06

P(will | positive) = $\frac{0+1}{5+12}$ = 0.06

P(back | positive) = $\frac{0+1}{5+12}$ = 0.06

P(dumb | positive) = $\frac{0+1}{5+12}$ = 0.06

<u>Secondly, for **neutral** sentiments:</u>

Total number of words in neutral sentences (n) = 3

Total number of words in vocabulary ($|Vocabulary|$) = 12

P(you | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07 P(don’t | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07

P(are | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07 P(like | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07

P(awesome | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07P(jealous | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07

P(thank | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07 P(will | neutral) = $\frac{1 +1}{3 + 12}$ = 0.13

P(be | neutral) = $\frac{1 +1}{3 + 12}$ = 0. 13 P(back | neutral) = $\frac{1 +1}{3 + 12}$ = 0.13

P(so | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07 P(dumb | neutral) = $\frac{0 +1}{3 + 12}$ = 0.07

Finally , for **neutral** sentiments:

Total number of words in negative sentences (n) = 10

Total number of words in vocabulary ($|Vocabulary|$) = 12

P(you | negative) = $\frac{3 +1}{10 + 12}$ = 0.18 P(don’t | negative) = $\frac{1 +1}{10 + 12}$ = 0.09

P(are | negative) = $\frac{2 +1}{10 + 12}$ = 0.14 P(like | negative) = $\frac{1 +1}{10 + 12}$ = 0.09

P(awesome | negative) = $\frac{0 +1}{10 + 12}$ = 0.05 P(jealous | negative) = $\frac{1 +1}{10 + 12}$ = 0.09

P(thank | negative) = $\frac{0 +1}{10 + 12}$ = 0.05 P(will | negative) = $\frac{0 +1}{10 + 12}$ = 0.05

P(be | negative) = $\frac{0 +1}{10 + 12}$ = 0. 05 P(back | negative) = $\frac{0 +1}{10 + 12}$ = 0.05

P(so | negative) = $\frac{1 +1}{10 + 12}$ = 0.09 P(dumb | negative) = $\frac{1 +1}{10 + 12}$ = 0.09

After training, the model can now determine whether the given context is positive, negative, or neutral based on the independent probability of words.

Text to classify: "You will be jealous"

$Y_{positive}$ = P(positive) * P(you | positive) * P(will | positive) * P(be | positive) * P(jealous | positive)

= $1.296e^{-5}$

$Y_{neutral}$ = P(neutral) * P(you | neutral) * P(will | neutral) * P(be | neutral) * P(jealous | neutral)

= $1.38e^{-5}$

$Y_{negative}$ = P(negative) * P(you | negative) * P(will | negative) * P(be | negative) * P(jealous | negative)

= $2.025e^{-5}$

As the probability of a sentence being negative has the most significant likelihood, it can be concluded that the given sentence has negative sentiment.

Overall, from the above section, the working mechanism of the Naïve Bayes theorem in NLP problems has been demonstrated. Moving further, the same algorithm, having a similar core concept, will be implemented to develop the sentiment analysis model.

3.3. Pseudocode for the solution

The pseudocode of the solution for the given problem domain is mentioned below:

```
IMPORT NumPy
IMPORT pandas
IMPORT training dataset
PRE-PROCESS dataset
SELECT features
CREATE vocabulary
CALCULATE frequency
CLASSIFY labels
TRAIN model
APPLY the Naïve Bayes Classifier
CALCULATE probabilities
FUNCTION predict_text:
    DO
        INPUT text to classify
        APPLY model
        CALCULATE positive probability
        CALCULATE negative probability
        CALCULATE neutral probability
        IF p(text | positive) > p(text | negative) AND p(text | positive) >
        p(text | neutral):
            RETURN positive
        ELSE IF $p(text \mid neutral) > p(text \mid negative)$ AND $p(text \mid neutral) > p(text \mid positive)$:
            RETURN neutral
        ELSE:
            RETURN negative
        END IF
    END DO
```

3.4. Flowchart of the solution

The diagrammatic representation of the solution for the problem domain is mentioned below:

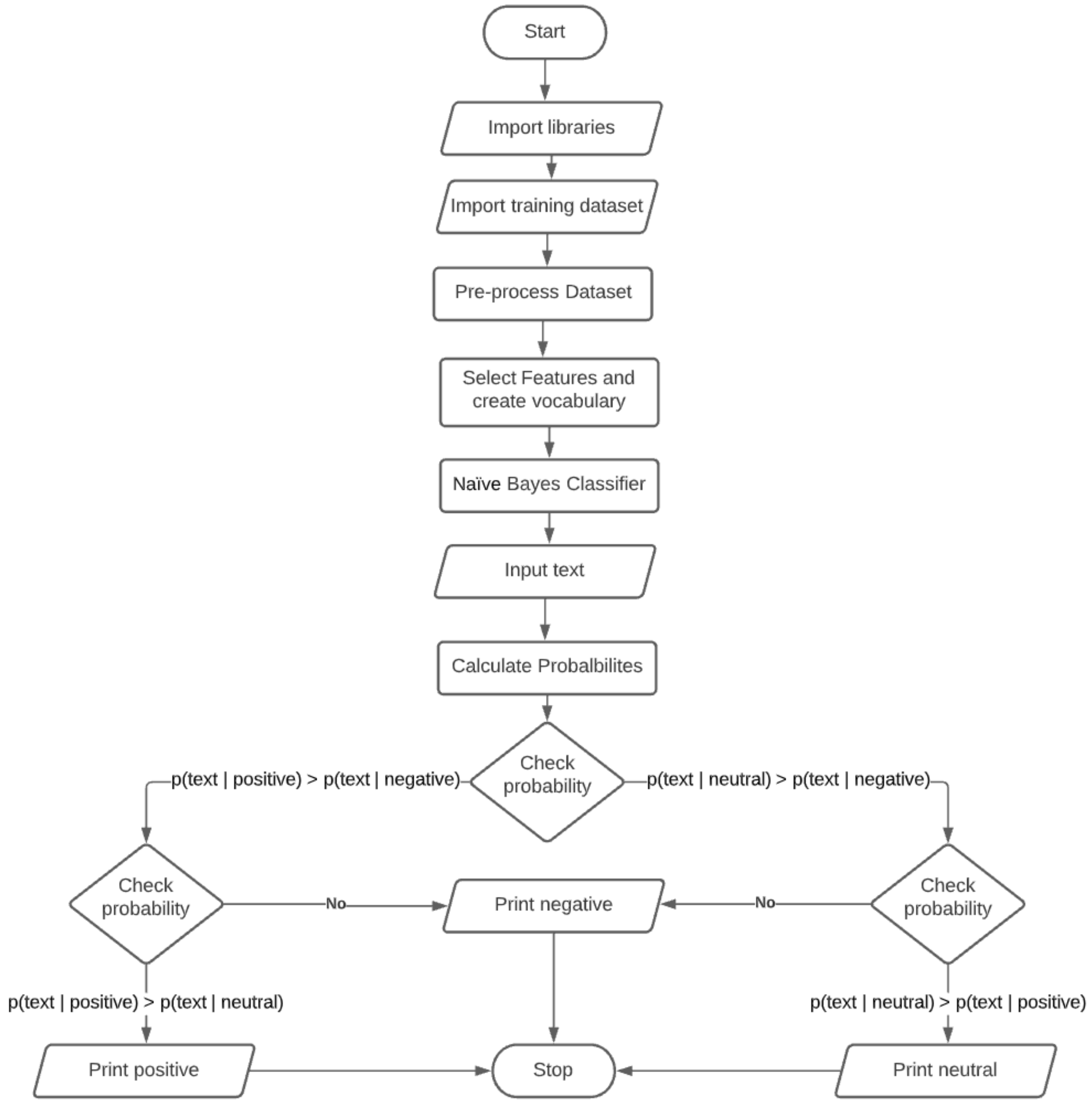

Figure 2: Flow chart

4. Conclusion

4.1. Analysis of the Work Done

Artificial Intelligence encompasses philosophy, psychology, and linguistics on top of computer science to develop intelligent systems. The task that generally requires human Intelligence can now be performed by intelligent agents (Duin & Bakhshi, 2017). Along with the rapid advancement in this field, the system that could imitate the simulation of human brains has been developed and concludes with promising results.

Overall, this project is based on the research and analysis of sentiment analysis that falls under the Natural Language Processing (NLP) subfield of AI. NLP, a fascinating but deeply intricated field, is concerned with synthesizing and analyzing written or spoken languages. This field focuses on developing applications that could effortlessly and efficiently distinguish human writings and speeches. But particularly in this project, a wide variety of algorithms, their limitations, and mitigating measures has been discussed for developing a sentiment analysis system.

The essentiality for developing an intelligent system that could detect human attitudes, emotions, and behavior towards services, products, or events is crucial for the remarkable growth of business and similar fields. The necessity for deploying such a system for mitigating probable incidents using enormous data being generated through the internet has been discussed in the project.

This report has thoroughly presented the intensive research based on sentiment analysis projects in AI and a proposed algorithm for developing similar applications. Various algorithms for solving sentimental analysis problems, including SVM, Naïve Bayes, CNN-LSTM, RNN, and more, have been portrayed. In concluding, the CNN-LSTM algorithm performs exceptionally well, resolving issues like bias and some extent of tone detection. But this algorithm requires high computational power and a long period of training. On the other hand, Naïve Bayes and SVM approach works fine within limited data and demands low computational power. For improving the performance of the Naïve Bayes algorithm, techniques like tokenization, lemmatization, stop words, stemming can be utilized. The transformation of words to vectors could gain higher accuracy in this model.

Overall, this project reflects abstracted research based on developing a sentiment analysis system. Including brief inclusion of Naïve Bayes theorem, review of similar projects, and explanation of problem domains, pseudocode, and flowchart of the proposed solution has been depicted in the documentation.

4.2. How the solution addresses the real-world problems

The thought and analysis of other people for making a decision has always been vital for every person. For instance, if one tends to purchase a laptop, the opinions of their friends or families will be taken into consideration. Further, they will perform intense research based on the content available on the internet in blogs, forums, and social networks. By figuring out the pros and cons of the product from the internet users' experience, one can finally make the purchasing decision.

By embracing ML algorithms to analyze customer sentiments, E-commerce companies can enhance their services and products based on users' experience. Also, companies will be facilitated in figuring out the present trends of markets. The application of sentiment analysis in the real world has a broader spectrum that incorporates highly demanded products, highly appreciated movies, mostly loved music, etc., by analyzing users' appraisals shared over the internet.

The field of sentiment analysis and opinion mining has also impacted brand analysis and political domains. With the help of sentiment analysis models, the government could accumulate the citizens' opinions upon the regulated policies. As a result, the guidelines could be upgraded and updated for the benefit of the people. In the current scenarios, big tech giants (companies) are hugely investing and adhering to the sentiment analysis approach to flourish their business strategies by understanding customers' opinions on their brands and products. On the top, few companies facilitate other companies to implement sentiment analysis for realizing trends of their services or products in public.

To conclude, the assumption is that the proposed model would be valuable for addressing the real-world issues that industries are striving to solve. Since this model focuses on implementing an independent probabilistic model, the analyzed opinions on specific topics will be highly accurate and effective to adopt.

4.3. Further work

After completing research, the explanation of the problem domain, analysis of similar projects and their outcomes, approach for solving similar problems have been included in this report. Further, a solution has been proposed for solving the sentiment analysis problem, which is based on the Naïve Bayes algorithm. The limitations of the proposed solution and the improvisation techniques for enhancing algorithm efficiency have been demonstrated. Finally, the pseudocode and flowchart have also been highlighted for evaluating the proper workflow of the implementation.

Advancing towards the project, implementing an algorithm for developing a sentiment analysis system is required to be accomplished. For which Jupyter Notebook, based on Python programming language, will be used to code for constructing the appropriate model. The relevant Python libraries like NumPy, pandas and more will be imported and used accordingly. The dataset discovered from the online resource must be pre-processed and imported into the system for training the model.

Now, the Naïve Bayes Classifier will be implemented for calculating independent probabilities and training the model from an imported dataset. Finally, the system will become ready to take input and classify the sentiment of text as either negative, positive, or neutral. By applying appropriate rules and techniques in the program file, the project will be completed producing refined system that can accurately analyze the core intent of the corpus.

References and Bibliography

Agarwal, B. & Mittal, N. (2016) *Prominent Feature Extraction for Sentiment Analysis*. 2nd ed. Stirling: Springer.

Ain, Q.T. et al. (2017) Sentiment Analysis Using Deep Learning Techniques: A Review. *International Journal of Advanced Computer Science and Applications*, 8(6), pp.424-33.

Alpaydin, E. (2016) *Machine learning : the new AI*. Cambridge: MIT Press.

Bird, S., Klein, E. & Loper, E. (2009) *Natural Language Processing with Python*. 1st ed. Sebastopol: y O'Reilly Media, Inc.

Boiy, E., Hens, P., Deschacht, K. & Moens, M.-F. (2007) Automatic Sentiment Analysis in On-line Text. In Chan, L. & Martens, B., eds. *11th International Conference on Electronic Publishing*. Vienna, 2007. IRIS-ISIS Publications.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013) New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), pp.15-21.

Camp, M.d. & Bosch, A.d. (2012) The socialist network. *Decision Support Systems*, 53(4), pp.761-69.

Deng, L. & Liu, Y. (2018) A Joint Introduction to Natural Language Processing and to Deep Learning. In L. Deng & Y. Liu, eds. *Deep Learning in Natural Language Processing*. Singapoore: © Springer Nature Singapore Pte Ltd. pp.1-23.

Duin, S.v. & Bakhshi, N. (2017) *Part 1: Artificial Intelligence Defined* [Online]. Available from: https://www2.deloitte.com/se/sv/pages/technology/articles/part1-artificial-intelligence-defined.html [Accessed 17 December 2021].

Gandhi, R. (2018) *Naive Bayes Classifier* [Online]. Available from: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [Accessed 15 December 2021].

Gupta, S. (2018) *Sentiment Analysis: Concept, Analysis and Applications* [Online]. Available from: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17 [Accessed 12 December 2021].

Hutto, C. & Gilbert, E. (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media /*, 8(1), pp.216-25.

Lane, P.C.R., Clarke, D. & Hender, P. (2012) On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems*, 53(4), pp.712-18.

Lee, K.C. (2021) *Sentiment Analysis — Comparing 3 Common Approaches: Naive Bayes, LSTM, and VADER* [Online]. Available from: https://towardsdatascience.com/sentiment-analysis-comparing-3-common-approaches-naive-bayes-lstm-and-vader-ab561f834f89 [Accessed 15 December 2021].

Lewis, T. (2014) *A Brief History of Artificial Intelligence* [Online]. Available from: https://www.livescience.com/49007-history-of-artificial-intelligence.html [Accessed 12 December 2021].

Liu, B. (2015) *Sentiment Analysis*. 1st ed. New York: Cambridge University Press.

Li, W. et al. (2020) User reviews: Sentiment analysis using lexicon integrated two-channel CNN-LSTM family models. *Applied Soft Computing Journal*, 94.

Maks, I. & Vossen, P. (2012) A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4), pp.680-88.

Medhat, W., Hassan, A. & Korashy, H. (2014) Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), pp.1093-113.

Mueller, J.P. & Massaron, L. (2021) *Machine Learning for Dummies*. 2nd ed. Hoboken: John Wiley & Sons, Inc.

Peterson, M. (2021) *Jan. 6 Was Worse Than We Knew* [Online]. Available from: https://www.nytimes.com/2021/10/02/opinion/jan-6-trump-eastman-election.html [Accessed 12 December 2021].

Ravi, K. & Ravi, V. (2015) survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89, pp.14-46.

Reddy, V. (2018) *Sentiment Analysis using SVM* [Online]. Available from: https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1 [Accessed 15 December 2021].

Rungta, K. (2018) *TensorFlow: Learn in 1 Day*.
Russell, S. & Norvig, P. (2010) *Artificial Intelligence A Modern Approach*. 3rd ed. New Jersey: Pearson Education, Inc.

Steinberger, J. et al. (2012) Creating sentiment dictionaries via triangulation. *Decision Support Systems*, 53(4), pp.689-94.

TechTarget. (2021) *sentiment analysis (opinion mining)* [Online]. Available from: https://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining [Accessed 12 December 2021].

Timberg, C., Dwoskin, E. & Albergotti, R. (2021) *Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs* [Online]. Available from:

https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/ [Accessed 12 December 2021].

Wang, J., Yu, L.-C., Lai, K.R. & Zhang, X. (2016) Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model. In *54th Annual Meeting of the Association for Computational Linguistics*. Berlin, 2016.