National College of Ireland

Configuration Manual

MSc Research Project
MSc Data Analytics

Venkata Ramya Bandaru
Student ID: 2215169

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. ……Bandaru Venkata Ramya……..……

Student ID: ……x22151699……..……..……

Programme: …MSc Data Analytics…… **Year:** ……2023……..

Module: …Research in Computing……..……

Lecturer: …Teerath Kumar Menghwar……..……

Submission Due Date: ……14/12/2023……..……

Project Title: …… Sentimental Analysis on Amazon Fine Food Reviews Using Machine Learning……

Word Count: ……1357…… **Page Count:** ……9……..……..……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……B. V. Ramya……

Date: 14/12/2023……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Venkata Ramya Bandaru
Student ID: 22151699

1 Introduction

This file contains configuration manual of the research work performed. Section 2 explains about system requirements, section 3 explains about hardware requirements, section 4 explains about libraries, section 5 explains about establishing database connection, section 6 explains about data gathering and visualization, section 7 explains about performing sentimental analysis, section 8 explains about results

2 System Requirements

Operating system- Windows 11

Ram – 8 GB

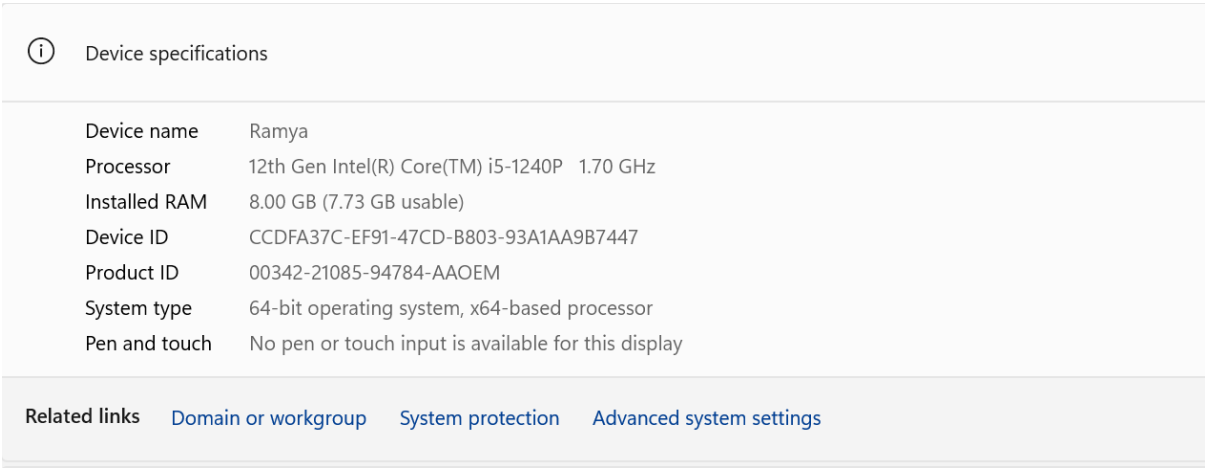
Processor – 11 Gen

Jupyter Notebook with version 6.5.4 to run the code

SQLite database to store the data

3 Hardware Requirements

System should atleast have 8 gb ram



4 Libraries needed

Numpy library should be installed and imported to perform numerical operations in python
Command - pip install numpy

Pandas library should be imported for data manipulation and analysis.
Command- pip install pandas

Matplotlib, seaborn library should be installed and imported to perform visualizations in python
Command - pip install matplotlib

SQLite3 should be installed for data storage.

Gensim library is used for document modelling and similarity analysis.
pip install genism

Time, random, warnings module should be imported.

Matplotlib Inline Magic sets the backend of matplotlib to inline for displaying plots.

Pickle module is used for saving and loading python objects to/from files. For this pickle module should be imported.

5 Establishing database connection

Connection to SQLite should be established and using read sql query and pandas the data is extracted from database and stored to data frame df1.

```
#Using sqlite3 to retrieve data from sqlite file

con = sql.connect("database.sqlite")#Connection object that represents the database

#Using pandas functions to query from sql table
df = pd.read_sql_query("""
SELECT * FROM Reviews
""",con)
```

6 Data Gathering and visualization

Amazon Fine Food Reviews dataset is collected from Kaggle

Link is provided : https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

Columns present in data are shown in fig1

```
df.head()
```

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Fig 1

Fig 2 shows the count of star rating in the dataset. Majority of the star ratings count is 5.

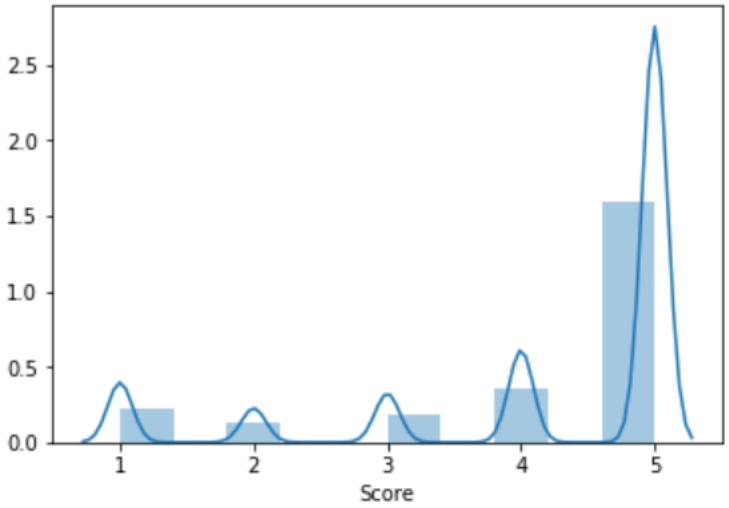


Fig 2

7 Sentimental Analysis

Firstly define polarity function. Then using the score column and map function the score column is labelled as positive, negative and neutral where neutral sentiments are ignored. This can be observed in fig 3 score column.

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	Positive	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	Negative	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	Positive	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	Negative	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	Positive	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Fig 3

As a part of data cleaning, remove duplicate rows by using drop duplicates function. Next remove HTML tags, Punctuations, stopwords.

Perform stemming and lemmatization on all the reviews.

After performing, stemming and lemmatization store all the positive and negative words in different variables which is further used to visualize most popular positive and negative words.

Fig 4, shows total count of positive and negative words in the dataset.

```
No. of positive words: 11678044
No. of negative words: 2393854

Most Common postive words [(b'not', 145019), (b'like', 138335), (b'tast', 126024), (b'good', 109838), (b'love', 106551), (b'fla
vor', 106408), (b'use', 102872), (b'great', 101125), (b'one', 94396), (b'product', 88466)]

Most Common negative words [(b'not', 53634), (b'tast', 33828), (b'like', 32059), (b'product', 27411), (b'one', 20176), (b'flavo
r', 18898), (b'would', 17858), (b'tri', 17515), (b'use', 15148), (b'good', 14616)]
```

Fig 4

WordCloud library should be downloaded to plot data in wordcloud image.

New data frame df2 is created where helpfulnessNumerator <= helpfulnessDenominator. Df2 have less rows when compared to original df as most of the rows got filtered based on the condition given.

Fig 5 shows the word cloud of positive words used in the reviews

Fig 5 wordcloud of positives reviews

7.1.1 Uni-gram BOW

Unigrams refer to a single word or token in a sequence of words.

Import countVectorizer from sklearn

Using this count vectorizer all the words in dataset df2 are converted to unigrams.

TruncatedSVD library should be imported to normalize the unigram vector data. After normalization almost 82.4 percent of the data is retrieved.

Before truncation and after truncation the data is saved to new files which can be used for further analysis

Datatime library is used to print date and time of the execution.

Now, the normalized unigram data is visualized with t-SNE(t-distributed stochastic neighbor embedding). Steps involved are
Setup seed for reproducibility
Configure tsne by taking only few data samples
Fit and transform the data using tsne and visualize the data in two dimensions.

TSNE model uses perplexity parameter which is used for dimentionality reduction and visualization of high dimensional data. Higher the perplexity higher will be the computational time.

Unigram visualizations with perplexity 30, 20, 40 are computed.

Unigram visualizations with perplexity 30 is shown in fig 6. Positive and negative words are displayed as facet grid in a two dimensional space

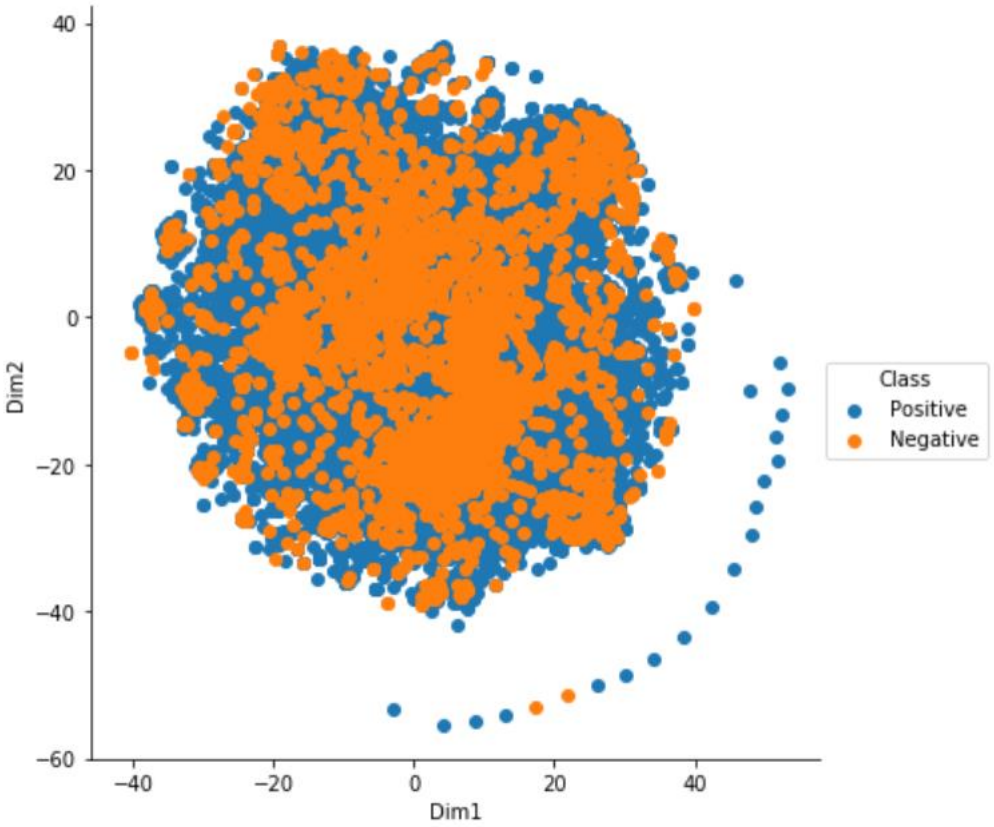


Fig 6

7.1.2 Bi-gram BOW

Bigrams refers to a two words or token in a sequence of words.

The same process that we performed for uni-gram bow is followed for Bi-gram BOW and by changing the perplexity values, facetgrids are plotted.

7.1.3 Tf-idf BOW

Term frequency which measures how often a term occurs in specific document and document frequency measures how many documents in a collection contain specific term and combination of tf and df evaluates the importance of a term.

Cleaned data from dataframe df2 is taken and weightages are assigned for all the words in the corpus.

TruncatedSVD library should be imported to normalize the tf-idf vector data.

The same process that we performed for uni-gram bow is followed for tf-idf vectorization and the data before and after performing normalization are stored in new files for further analysis.

By changing the perplexity values, facetgrids are plotted for tf-idf vectors to check the change in dimentionality reduction.

7.1.4 Word2Vec

To work with Word2Vec, Gensim module should be imported.

Gensim is a robust open-source vector space modeling and topic modeling toolkit implemented in Python. It uses NumPy, SciPy for performance. Gensim is specifically designed to handle large text collections, using data streaming and efficient incremental algorithms.

Word2Vec captures semantic relationships of data in continuous vector space.

The same process that we performed for uni-gram bow is followed for Word2Vec, Avg Word2Vec and hybrid approach of (TF-idf-W2Vec)

By changing the perplexity values, facetgrids are plotted to check the change in dimentionality reduction.

Most of the TSNE plot shows that data is quite overlapping hence we cant be sure that data is linearly separable. So Model training is done to determine the accuracy.

Explanation for "4 Amazon Food Reviews - Logistic Regression.ipynb" file

Import all the above mentioned libraries and from sklearn import accuracy_score, confusion matrix, precision score, f1_score, recall_score.

Original dataset after performing all the pre processing steps is splitted as train and test with size 70 and 30 percent.

Foor better results, using time series split the data is splitted into 10 splits using the library TimeSeriesSplit. This can be seen in fig 7

```
(23179, 161941) (23174, 161941)
(46353, 161941) (23174, 161941)
(69527, 161941) (23174, 161941)
(92701, 161941) (23174, 161941)
(115875, 161941) (23174, 161941)
(139049, 161941) (23174, 161941)
(162223, 161941) (23174, 161941)
(185397, 161941) (23174, 161941)
(208571, 161941) (23174, 161941)
(231745, 161941) (23174, 161941)
```

Fig 7

Model building using Logistic Regression

Create Logistic Regression Classifier model for data training and testing.

Change the hyper parameter value c between 0.001 to 1000 and use L1, L2 Regularizers.

Train and test the models to find the accuracy score of Unigrams, Bigrams, TF-IDF, Word2Vec, AvgWord2Vec, and hybrid(Tf-idf-Word2Vec) with both Grid SearchCV and Randomized SearchCV

8 Results

Logistic Regression (on whole dataset)					
Featurization	CV	Accuracy	F1-Score	C	Penalty
Uni - gram	GridSearch CV	91.95	0.749	10	L2
	Randomized Search CV	91.96	0.748	5	L2
Bi -gram	GridSearch CV	93.704	0.808	100	L2
	Randomized Search CV	93.704	0.808	100	L2
tfidf	GridSearch CV	93.615	0.809	5	L1
	Randomized Search CV	93.616	0.809	5	L1
Avg Word2Vec	GridSearch CV	89.28	0.638	1000	L2
	Randomized Search CV	89.264	0.637	10	L2
tfidf - Word2vec	GridSearch CV	88.027	0.535	10	L2
	Randomized Search CV	88.093	0.531	5	L2

Bigram Featurization performs best with accuracy of 93.704 and F1-Score of 0.808 • Sparsity increases as we increase lambda or decrease C when L1 Regularizer is used.