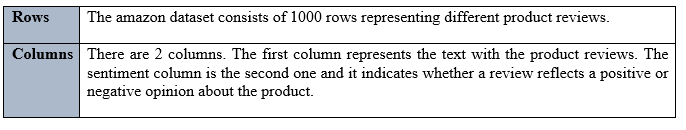
**CLASSIFICATION AND TOPIC DETECTION ON AMAZON REVIEWS**

**TASK 1: TEXTUAL DATA PREPROCESSING OF THE AMAZON DATASET**

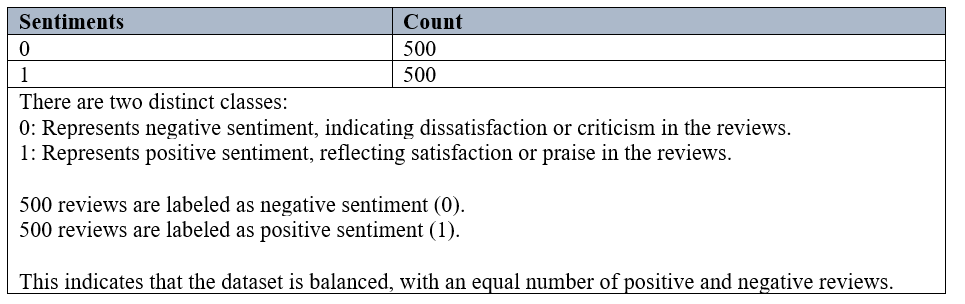
**1.1 DATA OVERVIEW**

For this project, I used the Amazon dataset containing customer product reviews.

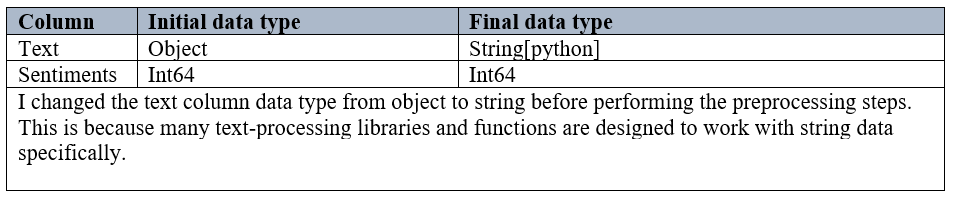
**1.1.1 Data Shape**



**1.1.2 Sentiment counts**



**1.1.3 Data Types**



**1.2 TEXTUAL DATA PREPROCESSING**

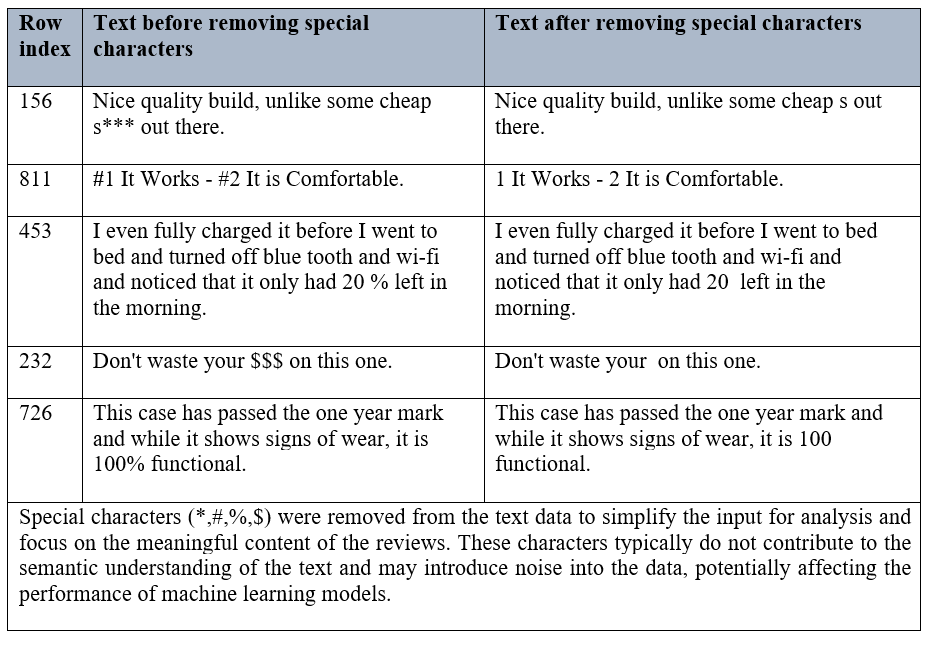
This is the data preprocessing pipeline that I followed to prepare the data for modelling.



**1.2.1 Text Extraction**

|  |
| --- |
|  |
| In this step, the reviews from the dataset were extracted from the Text column for preprocessing. The Text column contains the written reviews provided by customers, which are the primary source of data for text analysis. By isolating this column into a variable named text, the subsequent preprocessing steps can be applied efficiently. |

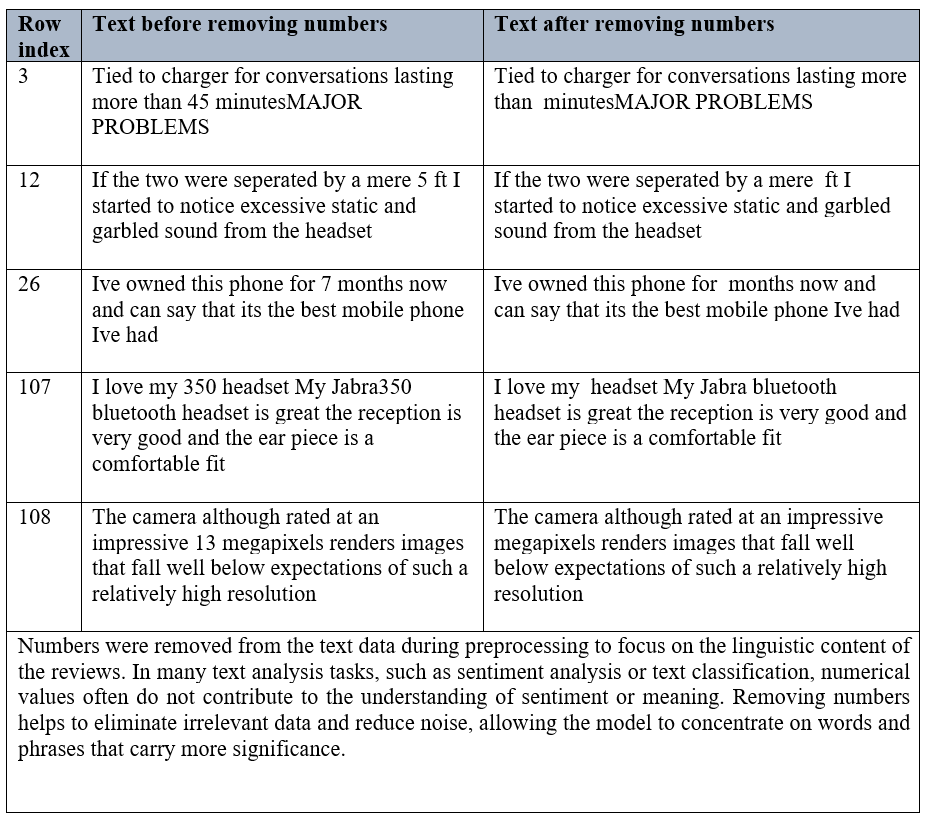
**1.2.2 Removing Special Characters**



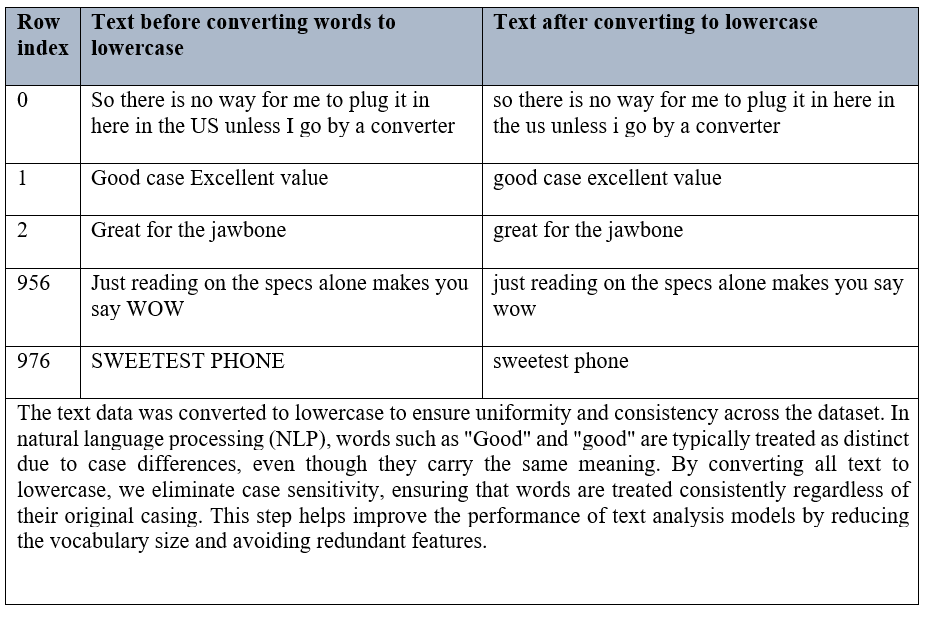
**1.2.3 Removing Punctuation**

|  |  |  |
| --- | --- | --- |
| **Row index** | **Text before removing punctuation** | **Text after removing punctuation** |
| 1 | Good case, Excellent value. | Good case Excellent value |
| 3 | Tied to charger for conversations lasting more than 45 minutes.MAJOR  PROBLEMS!! | Tied to charger for conversations lasting more than 45 minutesMAJOR PROBLEMS |
| 7 | If you are Razr owner...you must have this! | If you are Razr owneryou must have this |
| 8 | Needless to say, I wasted my money. | Needless to say I wasted my money |
| 25 | Great Pocket PC / phone combination. | Great Pocket PC phone combination |
| Punctuation marks (e.g., !?) were removed from the text data as part of preprocessing to streamline the analysis. While punctuation can sometimes convey tone or emphasis, it is generally not essential for understanding the overall meaning of the text in tasks like sentiment analysis or classification. Removing punctuation helps reduce noise in the dataset and ensures a more consistent and uniform representation of the text. This step is particularly useful when applying tokenization or vectorization techniques, as it focuses on the words themselves rather than symbols. | | |

**1.2.4 Removing Numbers**



**1.2.5 Converting words to lowercase**



**1.2.6 Removing Stop Words**

|  |  |  |
| --- | --- | --- |
| **Row index** | **Text before removing stop words** | **Text after removing stop words** |
| 10 | and the sound quality is great | sound quality great |
| 20 | i went on motorolas website and followed all directions but could not get it to pair again | went motorolas website followed directions could get pair |
| 30 | this is a simple little phone to use but the breakage is unacceptible | simple little phone use breakage unacceptable |
| 40 | it has a great camera thats mp and the pics are nice and clear with great picture quality | great camera thats mp pics nice clear great picture quality |
| 50 | not loud enough and doesnt turn on like it should | loud enough doesnt turn like |
| Stop words (e.g., "the", "is", "in", "and", "to") were removed from the text data during preprocessing to improve the efficiency and effectiveness of the analysis. Stop words are common words that occur frequently in language but carry little meaningful information in the context of text analysis. These words often do not contribute to the sentiment or key themes of the text and can add unnecessary noise to the dataset. By removing stop words, we reduce the dimensionality of the data, allowing the model to focus on more informative words that are likely to be more relevant for tasks like sentiment analysis or classification. | | |

**1.2.7 Lemmatizing**

|  |  |  |
| --- | --- | --- |
| **Row index** | **Text before lemmatizing** | **Text after lemmatizing** |
| 100 | integrated seamlessly motorola razr phone | integrate seamlessly motorola razr phone |
| 120 | ive tried several different earpieces cell phone jabra one first one ive found fits ear comfortably | ive try several different earpiece cell phone jabra one first one ive found fit ear comfortably |
| 150 | hoping | hop |
| 210 | bother contacting company dollar product learned lesson bought form online anyway | bother contact company dollar product learn lesson bought form online anyway |
| 220 | freezes frequently | freeze frequently |
| Lemmatization was applied to the text data to reduce words to their base or root form. Unlike stemming, which cuts off word endings to produce a root form, lemmatization considers the context and converts words to their correct base form (e.g., "running" to "run," "better" to "good"). This process helps to standardize words with different inflections, ensuring that variations of the same word are treated as a single feature. By lemmatizing the text, we reduce redundancy in the dataset and improve the quality of the input for machine learning models, as it focuses on the core meaning of words rather than their specific forms. | | |

**TASK 2: CLASSIFICATION**

**2.1 DATASET PREPARATION**

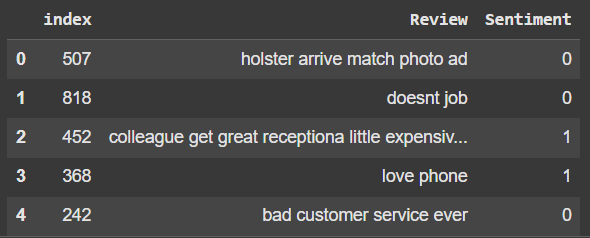
**2.1.1 Data reshuffling**

Before performing classification, I reshuffled the order of the text to prevent any order from affecting the performance.

Text order before reshuffling:

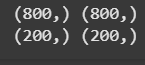


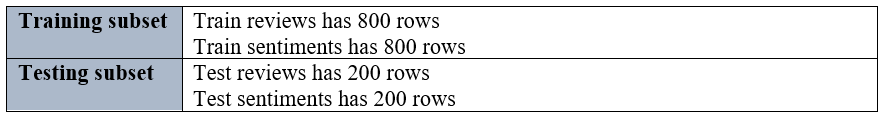
Text order after reshuffling:



**2.1.2 Dataset splitting**

I then split the dataset into training and test sets and this was the output:

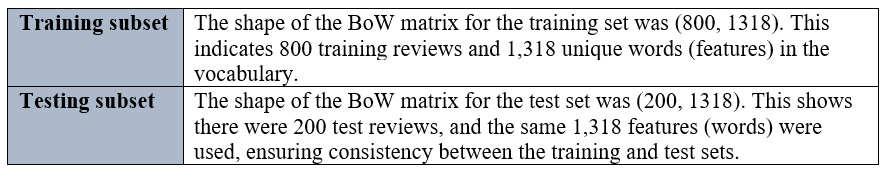




**2.1.3 Bag of Words Representation Using Count Vectorizer**

The Bag of Words (BoW) approach was used to convert textual data into a numerical format suitable for machine learning models. This was implemented using the CountVectorizer class from the sklearn.feature\_extraction.text module.

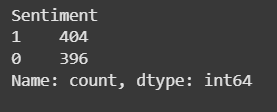
This was the output:  

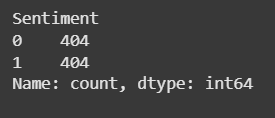
**2.1.4 Balancing the classes**

I used the RandomOverSampler technique to address the imbalance in the classes. This technique works by randomly duplicating instances from the minority class to balance the number of instances between the classes. It is applied during the training phase, ensuring that the model is trained on a balanced dataset.

Class distribution before oversampling:



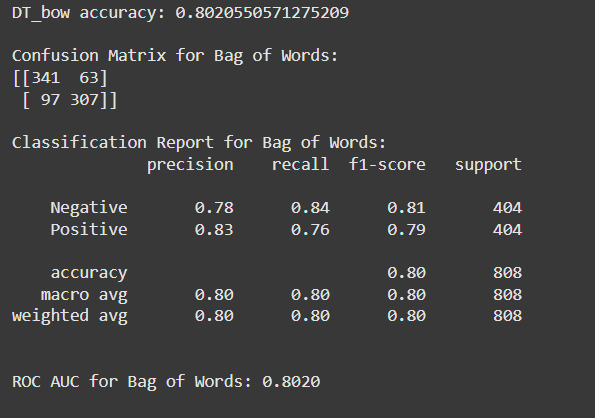
Class distribution after oversampling:



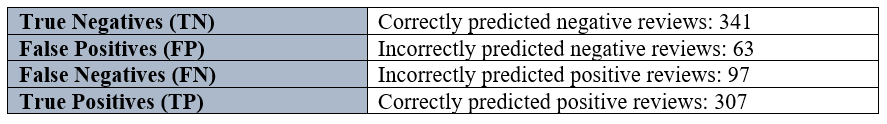
I used the following algorithms to perform classification on the amazon dataset:

1. Decision Tree
2. K-Nearest Neighbor
3. Naïve Bayes
4. Random Forest

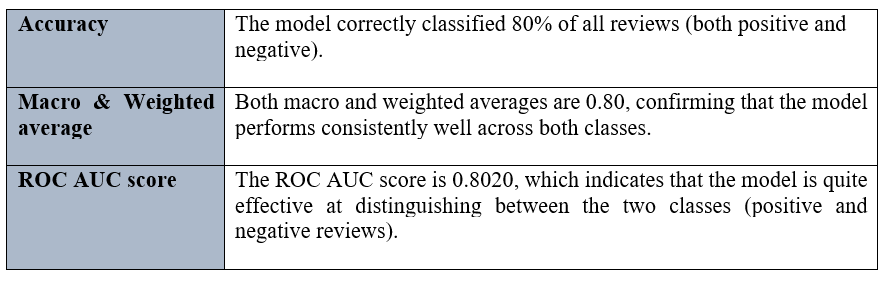
**2.2 DECISION TREE CLASSIFICATION**



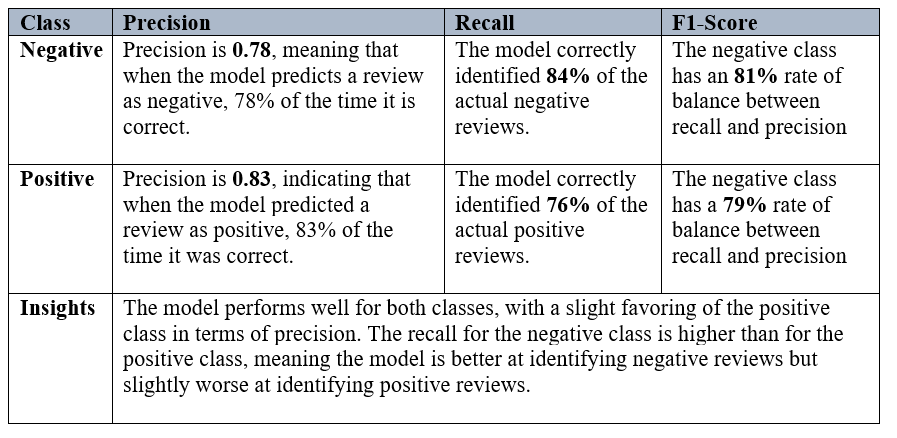
**Confusion Matrix**



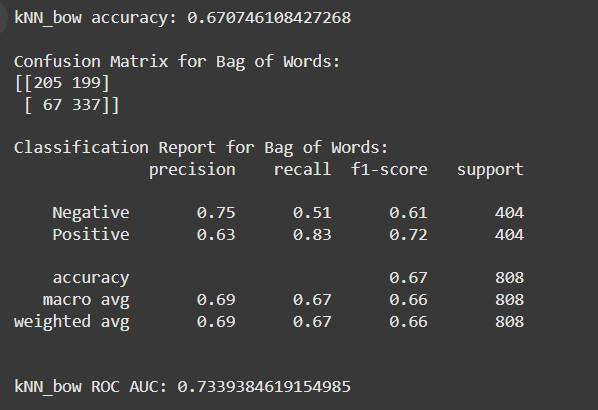
**Overall Model Performance**



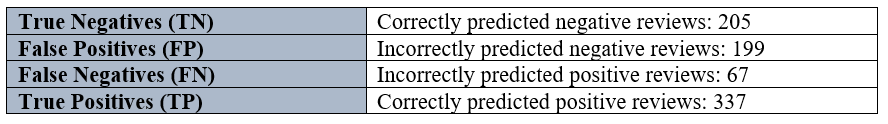
**Decision Tree Class Metrics Analysis**



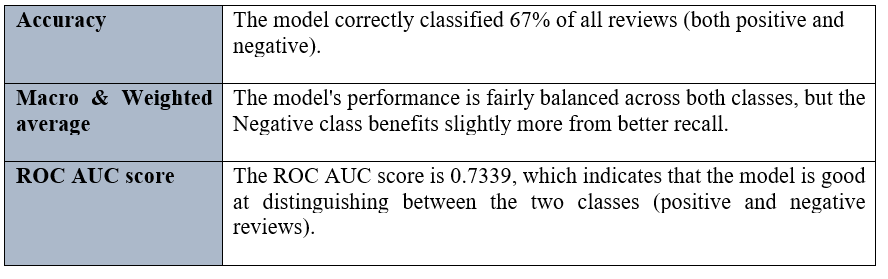
**2.3 k-NEAREST NEIGHBOUR**



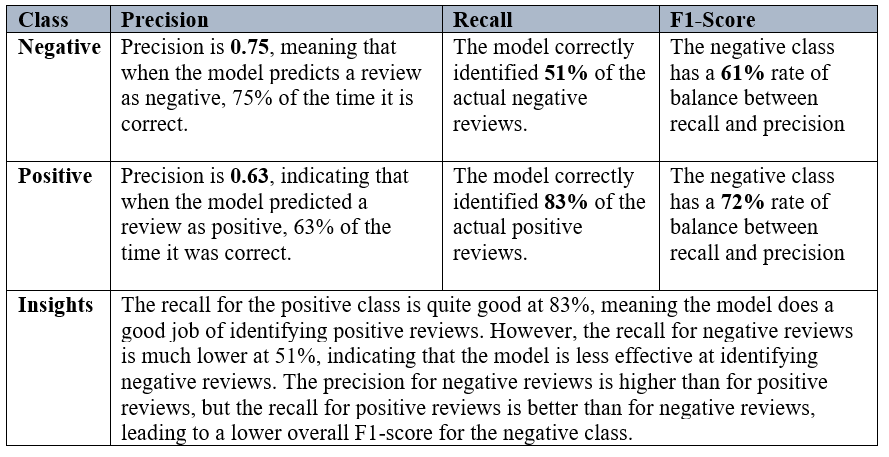
**Confusion Matrix**



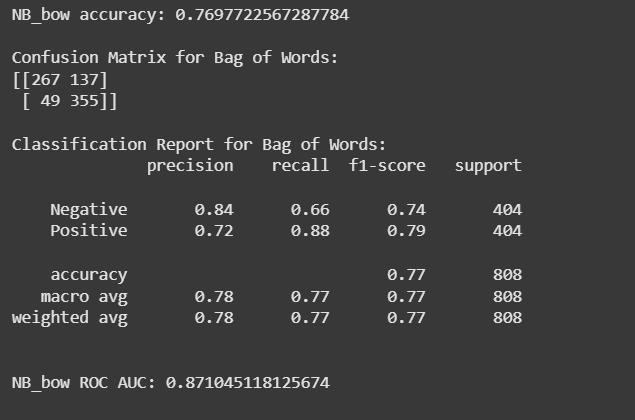
**Overall Model Performance**



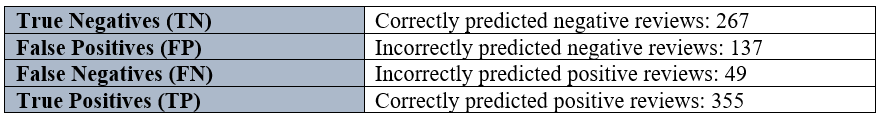
**kNN Class Metrics Analysis**



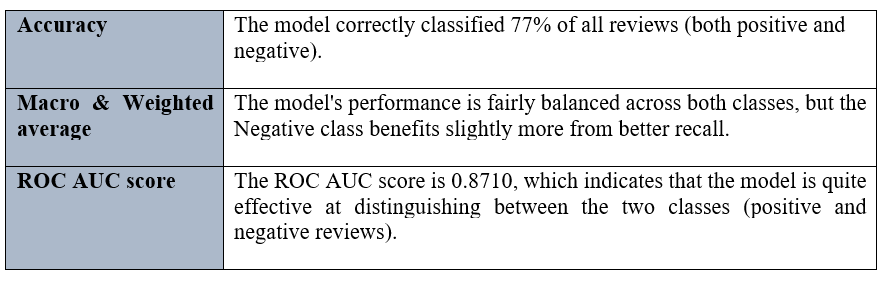
**2.4 NAÏVE BAYES**



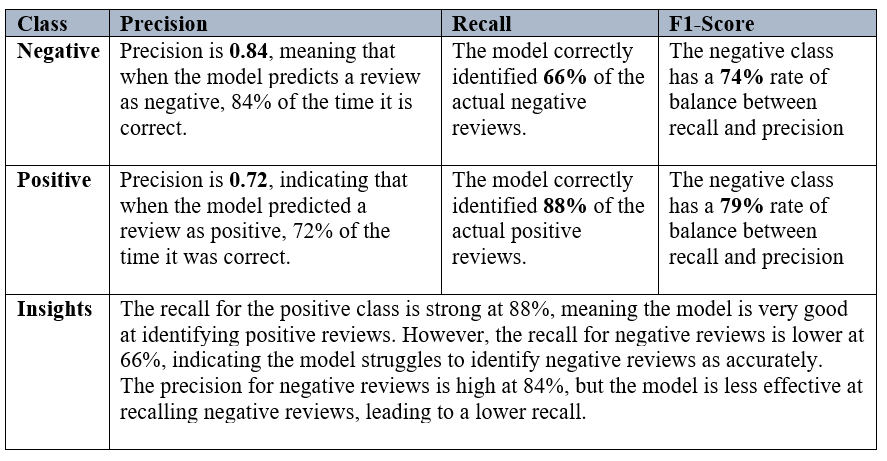
**Confusion Matrix**



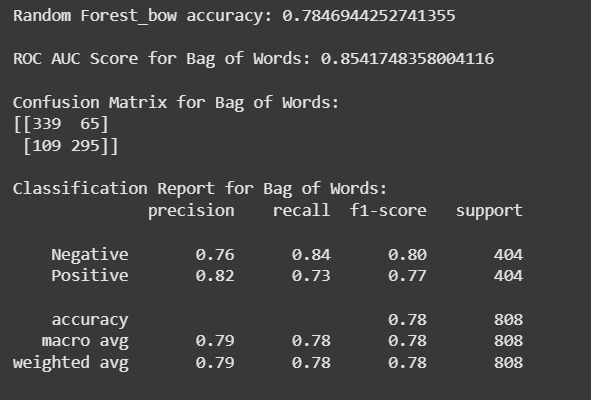
**Overall Model Performance**



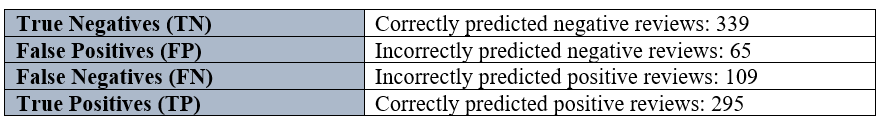
**Naïve Bayes Class Metrics Analysis**



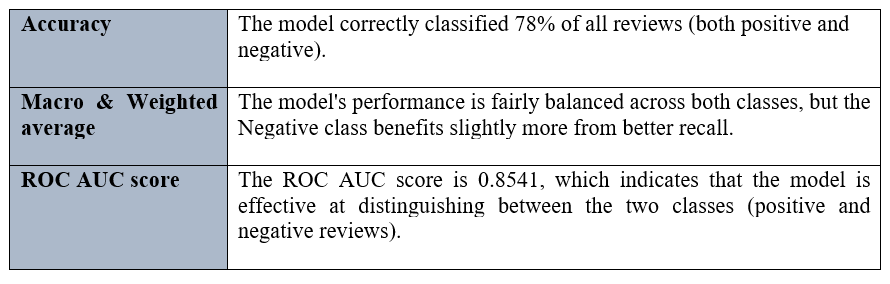
**2.5 RANDOM FOREST**



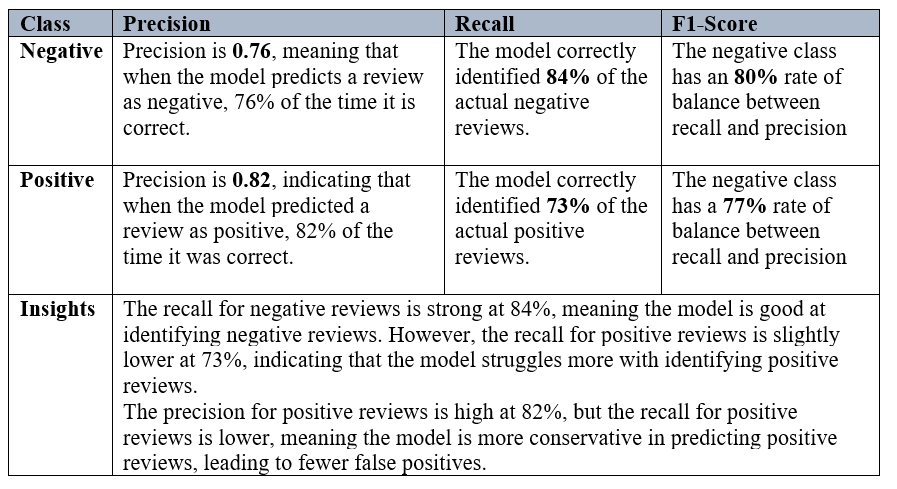
**Confusion Matrix**



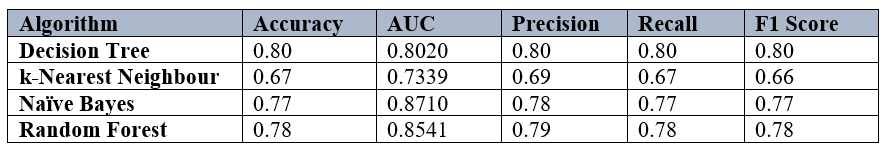
**Overall Model Performance**



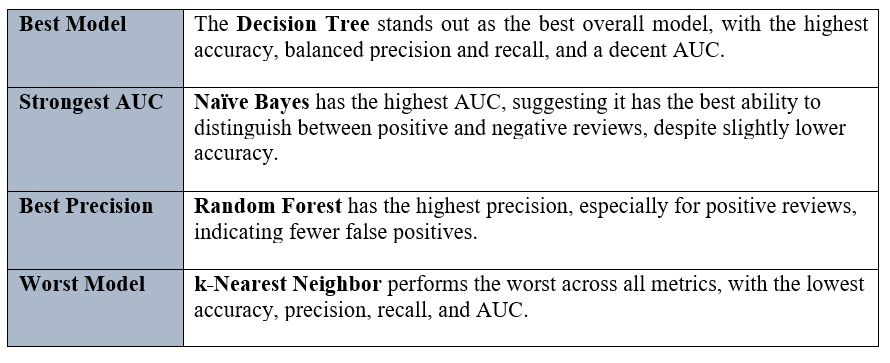
**Random Forest Class Metrics Analysis**



**2.6 COMPARISON OF THE CLASSIFICATION ALGORITHMS**



**Conclusion**



**TASK 3: BERT CLASSIFICATION**

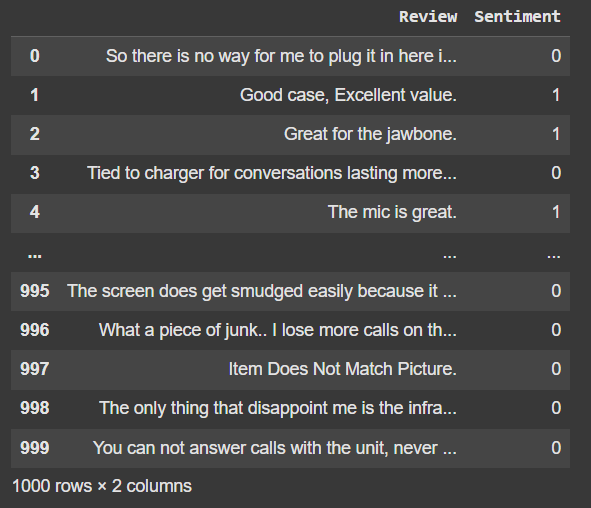
I performed classification on the amazon dataset using a BERT-based model with fine-tuning.

**3.1 DATASET PREPARATION**

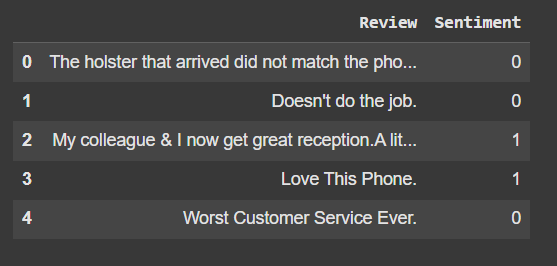
**3.1.1 Data reshuffling**

Before performing BERT classification, I reshuffled the order of the text to prevent any order from affecting the performance.

Text order before reshuffling:

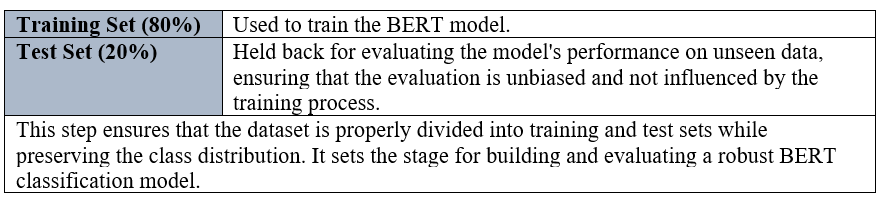


Text order after reshuffling:

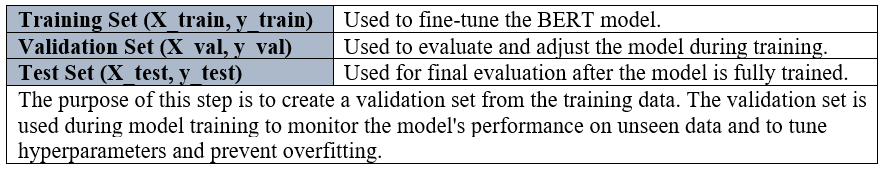


**3.1.2 Dataset Splitting**

I split the dataset into two parts:



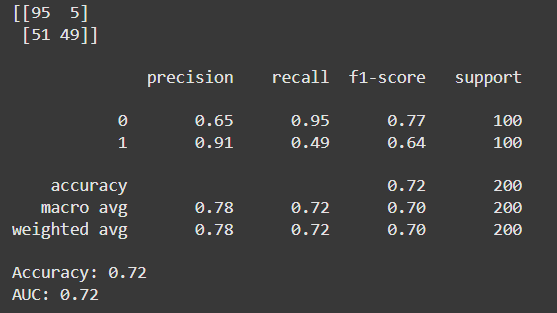
The training set (X and y) was further split into a training subset and a validation subset.



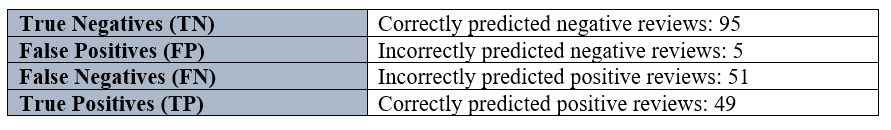
**3.2 BERT MODEL SELECTION**

|  |  |
| --- | --- |
| **BERT Model Selected** | URL: https://tfhub.dev/tensorflow/small\_bert/bert\_en\_uncased\_L-4\_H-512\_A-8/1  This URL points to the TensorFlow Hub resource for the selected pre-trained BERT model:  Small BERT Variant: A lightweight version of BERT with:  4 layers (L-4).  512 hidden size (H-512).  8 attention heads (A-8).  This model is pre-trained and optimized for tasks requiring efficient computation while retaining good performance. |
| **Preprocess Model Auto-Selected** | URL: <https://tfhub.dev/tensorflow/bert_en_uncased_preprocess/3>  This URL points to the preprocessing module that prepares input text for the selected BERT model:  Converts raw text into tokenized input.  Adds special tokens ([CLS] and [SEP]).  Truncates or pads sequences to match the model's input size. |

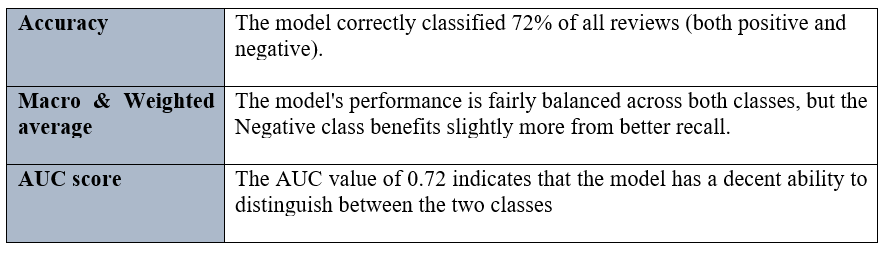
**3.3 BERT Classification results**



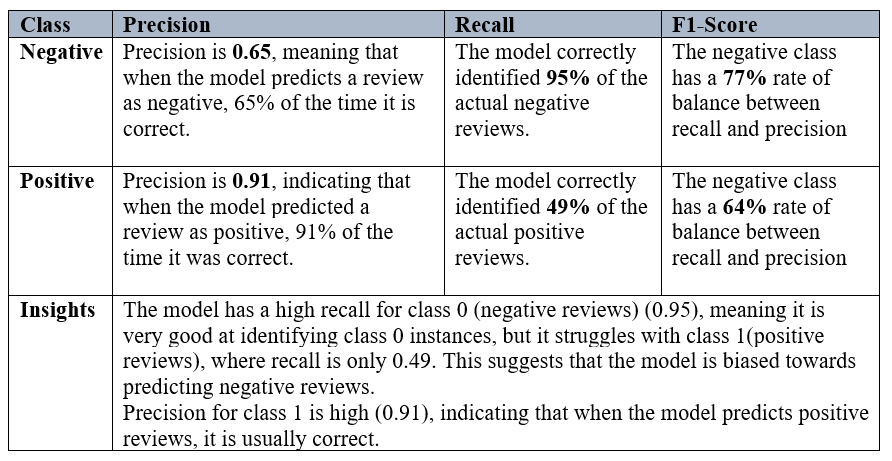
**Confusion Matrix**



**Overall Model Performance**

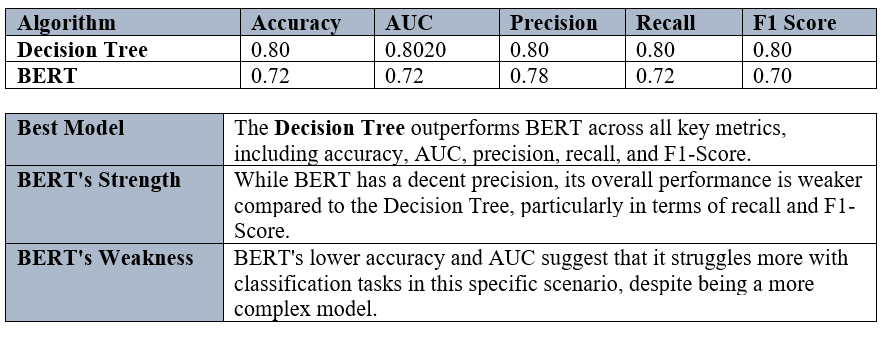


**BERT Class Metrics Analysis**

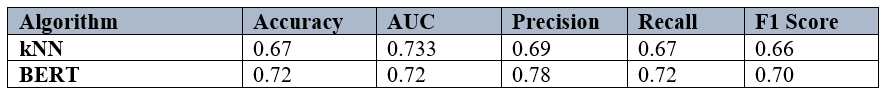


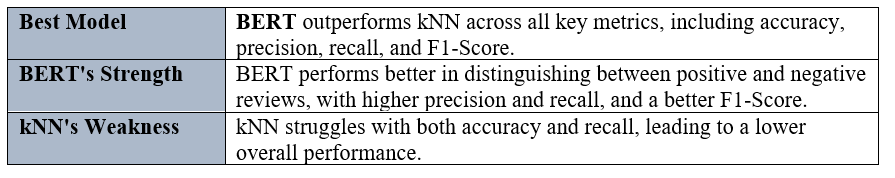
**3.4 MODEL COMPARISON BETWEEN BERT CLASSIFICATION AND OTHER CLASSIFICATION ALGORITHMS**

**3.4.1 BERT vs Decision Tree Classification**

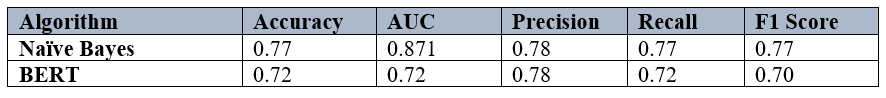


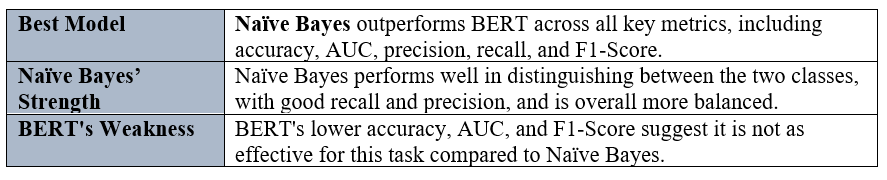
**3.4.2 BERT vs k-Nearest Neighbour**



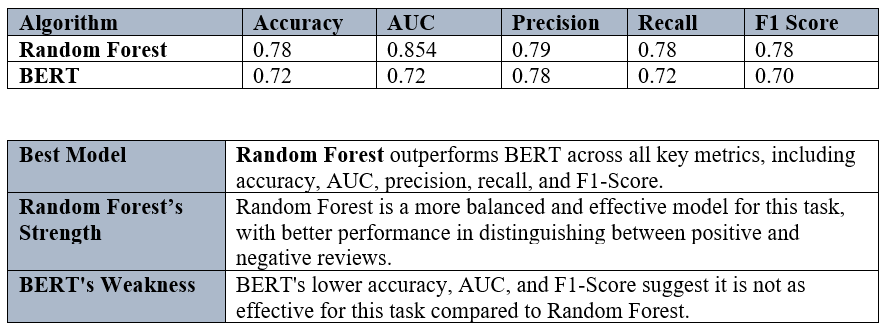


**3.4.3 BERT vs Naïve Bayes**





**3.4.4 BERT vs Random Forest**



**Conclusion**

In this comparison, BERT outperforms k-Nearest Neighbour (kNN) but falls behind Decision Tree, Naïve Bayes, and Random Forest regarding accuracy, AUC, precision, recall, and F1-Score.

**TASK 4: TOPIC DETECTION**

I used the Latent Dirichlet Allocation (LDA) algorithm and the Non-Negative Matrix Factorization (NMF) algorithm to detect topics in the Amazon reviews dataset.

**4.1 LATENT DIRICHLET ALLOCATION (LDA)**

I generated the following topics with 30 words each from the LDA algorithm.

|  |  |
| --- | --- |
| **Topics** | **Interpretation** |
| **1: phone, excellent, well, recommend, product, sound, quality, work, headset, price, battery, time, cheap, bad, ive, low, highly, really, would, case, make, good, like, around, one, beep, come, great, car, fit** | This topic concerns headsets or phone accessories with feedback regarding sound quality, pricing, and battery life. |
| **2:** **phone, work, get, great, like, use, product, don’t, time, best, one, service, buy, headset, well, make, plug, long, poor, doesn’t, new, ive, customer, volume, last, thing, easy, bad, really battery** | This topic includes mixed feedback on accessories like chargers, headphones, or phone cases. Some users appreciate their utility, while others face issues like poor performance or volume |
| **3:** **phone, work, go, love, use, fit, even, well, nice, charge, look, product, headset, get, great, day, try, also, like, service, bad, one, im, cell, case, cool, return, say, call, good** | This topic focuses on phone cases or similar products like headsets. Users appreciate their fit, aesthetic design, and overall appearance. |
| **4: phone, great, good, work, headset, don’t, battery, case, really, life, say, ive, drop, charger, im, sound, get, light, Motorola, long, Bluetooth, car, best, time, use, waste, money, that’s, new, product** | This topic focuses on phone batteries, Bluetooth headsets, and Motorola products. Users have reviewed the sound quality and price value. |
| **5:** **great, battery, product, phone, work, would, recommend, good, make, charger, price, well, im, sound, get, quality, use, car, purchase, item, simple, picture, bought, even, go, bad, problem, anyone, horrible, life** | This topic revolves around car chargers, phone battery life, and Bluetooth adapters. Users express mixed opinions about the quality and functionality. Some users would recommend these products to others. |
| **6: use, sound, headset, quality, good, great, phone, time, ear, couldn’t, like, service, really, drop, work, end, could, comfortable, charge, call, look, bad. Disappoint, put, loud, case, hour, feature, need, take** | This topic relates to headsets. Users are disappointed with comfort, sound quality, or the product's inability to perform as expected (e.g., dropping calls or being too loud). |
| **7: work, fine, phone, use, headset, well, call, product, battery, quality, happy, camera, try, time, case, great, purchase, fit, im, usb, avoid, charm, belt, company, broke, voice, much, right, get, best** | This topic is likely about smartphone accessories like headsets and USB cables. Users are happy with the overall battery quality, camera performance, and company support. |
| **8: use, well, quality, get, good, phone, product, make, work, ear, don’t, software, screen, excellent, reception, could, first, love, case, sound, small, two, bought, money, send, didn’t, earpiece, low, look, call** | This topic focuses on software or network reception issues with smartphones or accessories. While some users praise the product's quality, others face problems like poor software functionality or bad reception. |
| **9: great, phone, work, one, ear, problem, good, button, also, price, comfortable, easy, charge, sound, well, item, give, hold, thing, look, quality, piece, try, ive, accidentally, get, turn, im, mention, technology** | This topic highlights user satisfaction with the comfort and functionality of headphones or earbuds. Buttons and controls seem intuitive and easy to use. |
| **10: good, ear, great, nice, headset, make, phone, case, would, seller, year, money, purchase, problem, price, battery, one, contact, several, still, plug, first, go, excellent, look, last, also, screen, company, buy** | This topic addresses the buying process, customer service, and overall satisfaction. While some users report a smooth experience and product quality, others highlight issues with sellers or receiving defective items. |

**4.2 NON-NEGATIVE MATRIX FACTORIZATION (NMF)**

Non-Negative Matrix Factorization (NMF) is a dimensionality reduction technique widely used for topic detection in textual data. In the context of text analysis, NMF is applied to the TF-IDF matrix, where rows represent documents, and columns represent terms. The algorithm iteratively optimizes the matrices to minimize the reconstruction error while ensuring all values remain non-negative.

This is the data preprocessing pipeline that I used before performing topic detection with the NMF algorithm:



I generated the following topics with 30 words each from the LDA algorithm.

|  |  |
| --- | --- |
| **Topics** | **Interpretation** |
| **1: great, price, deal, item, device, worked, working, little, charger, earpiece, jawbone, absolutely, mic, service, choice, value, sound, reception, using, software, audio, phone, thank, headphones, expensive, nice, bluetooth, case, plug, motorolas** | This topic likely represents reviews about Motorola audio products like headphones. The users are discussing product performance, pricing, and value for money, including positive mentions of deals or complaints about expensive items. |
| **2: good, price, case, far, really, audio, quality, amazon, disappointing, real, reception, design, definitely, value, protection, bargain, think, fine, work, sunglasses, camera, pretty, seller, samsung, purchase, looks, volume, stuff, nice, reasonably** | This topic likely represents reviews about sunglasses, cameras, and Samsung phones and products. Reviews focus on product quality, phone reception, and design, and whether the item is worth the price. Mixed sentiments could be inferred. |
| **3: phone, ive, worst, best, new, does, like, work, cell, doesnt, im, did, sturdy, completely, buy, cool, mobile, bought, nokia, store, beautiful, car, got, say, horrible, dropped, purchase, camera, hear, charm** | This topic likely represents feedback on Nokia mobile phone performance and camera quality, with contrasting opinions about "best" and "worst," suggesting a mix of positive and negative reviews. |
| **4: product, excellent, price, happy, buy, dont, worthless, use, really, big, impressed, high, ordered, company, pretty, purchase, easy, satisfied, difficult, like, review, set, unhappy, using, exactly, fails, going, amazon, sure, described** | This topic likely represents general satisfaction or dissatisfaction with the product, emphasizing both positive and negative experiences. |
| **5: works, fine, just, plug, like, comfortable, packaged, charm, lightweight, work, described, arrived, new, advertised, time, better, quickly, car, microphone, lot, device, priced, handsfree, high, people, right, shipped, perfectly, fantastic, company** | This topic likely represents feedback about hands-free microphones. The reviews highlight whether it works as advertised, how it was shipped and packaged, and its usability, with a focus on practical aspects like. |
| **6: quality, sound, poor, service, low, nice, audio, clear, excellent, better, time, happy, bad, crap, talk, cheap, stars, end, terrible, customer, voice, ear, camera, hear, construction, clarity, poorly, regarding, people, design** | This topic likely represents feedback on sound quality (e.g., clarity or poor audio) and customer service experiences, with polarizing opinions. |
| **7: waste, money, dont, time, buy, make, mistake, want, terrible, bought, did, wasted, nokia, big, unit, like, just, basically, thank, work, right, support, beware, doesnt, return, expect, problems, usually, fails, say** | This topic likely represents feedback on Nokia products. Negative reviews express frustration over wasted money, faulty products, or poor performance. |
| **8: recommend, highly, item, case, really, nice, device, impressed, difficult, priced, comfort, try, looks, overall, blue, cool, tooth, new, leather, people, bad, buying, wouldnt, amazon, unit, comes, jabra, family, different, igo** | This topic likely represents feedback on Jabra products. The reviews have recommendations for specific products, emphasizing aesthetics or unique features, though some reviews note difficulties. |
| **9: headset, love, excellent, bluetooth, best, ear, comfortable, reception, just, use, used, thing, really, ive, jabra, fit, impressed, bt, using, pleased, fits, time, sound, wireless, plantronics, purchase, market, logitech, easy, right** | Reviews about Jabra headsets and Bluetooth devices, often with positive mentions of comfort and functionality. |
| **10: battery, life, disappointed, long, service, terrible, bought, buy, use, dying, original, problem, completely, quickly, extended, vi, holding, unreliable, new, complaint, useless, right, hours, excellent, quite, junk, im, motorola, real, highly** | This topic likely represents feedback on Motorola products and accessories. The reviews are focusing on battery performance, with complaints about short battery life or unreliable functionality. |

**4.3 MODEL COMPARISON**

| **Aspect** | **LDA** | **NMF** |
| --- | --- | --- |
| **Interpretability** | Broader, overlapping topics | Distinct, focused topics |
| **Sentiment** | Mixed within topics | Separated positive and negative sentiments |
| **Product Features** | General themes | Specific features like *battery* or *Bluetooth* |
| **Overlap** | High overlap between topics | Low overlap, more distinct topics |
| **Actionable Insights** | Requires further interpretation | Easier to derive actionable insights |

**Conclusion**

* **NMF** provides more distinct and actionable topics, making it better suited for deriving specific product insights and addressing customer concerns.
* **LDA** offers a broader view of customer feedback but requires more effort to interpret due to overlapping themes and mixed sentiments.