Assessment Submission Cover Sheet

This Assessment Cover Sheet **must** be included on all Assessment submissions.

|  |  |
| --- | --- |
| Assignment Title | Text Mining with Sentiment Analysis |
| Module | Data Mining |
| Student Name  (same as Student Card) | Jonas Wortmann |
| Student Number | D213125503 |
| Programme | TU059 |
| Part-Time/Full-Time | Full-Time |
| Year of Study  (First Year, Second Year, etc) | First Year |

Late Submissions: Assessment submitted after the deadline will have a late penalty applied.

**Academic Integrity for assessment in TU Dublin Programmes**

Each student is responsible for knowing and abiding by TU Dublin Academic Regulations and Policies. Any student in breach of these regulation/policies will be subject to action in accordance with the University’s procedures for breaches of assessment regulations. Please refer to the General Assessment Regulations at

<https://tudublin.libguides.com/c.php?g=674049&p=4794713>

<https://www.tudublinsu.ie/advice/exams/breachesofregulations/>

All students are expected to complete their courses/programmes in compliance with University regulations. No student shall engage in any activity that involves attempting to receive a grade by means other than honest effort, for example:

1. No student shall complete, in part or in total, any examination or assessment for another person.
2. No student shall knowingly allow any examination or assessment to be completed, in part or in total, for themselves by another person.
3. No student shall plagiarise or copy the work of another and submit it as their own work.
4. No student shall falsify any data. Falsification is the invention of data, its alteration, its copying from any other source, or otherwise obtaining it by unfair means, or inventing quotations and/or references.
5. No student shall use aids or devices excluded by the lecturer in undertaking course work or assessments/ examinations.
6. No student shall knowingly procure, provide, or accept any materials that contain questions or answers to any examination or assessment to be given at a subsequent time.
7. No student shall provide their assignments, in part or in total, to any other student in current or future classes of this module/ programme unless authorised to do so by the lecturer.
8. No student shall submit substantially the same material in more than one module/programme without prior authorization.
9. No student shall alter graded assignments or examinations and then resubmit them for regrading, unless specifically authorised to do so by the lecturer.
10. All programming code and documentation, unless correctly referenced, submitted for assessment or existing in the student’s computer accounts must be the students’ original work or material specifically authorized by the lecturer.
11. Collaborating with other students to develop, complete or correct course work is limited to activities explicitly authorized by the lecturer.
12. For all group assignments, each member of the group is responsible for the academic integrity of the entire submission. Consequently, all group members must satisfy themselves that all elements of their submission adhere to the academic integrity statement points above.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University’s Regulations and Policies covering Academic Integrity (see General Assessment Regulations)*.*

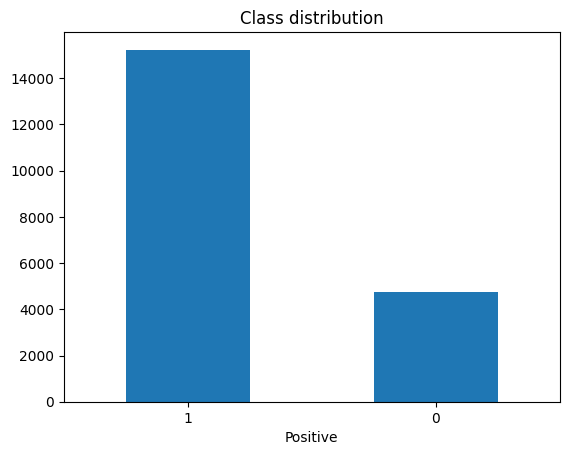
Coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. If you have queries about what constitutes plagiarism, please speak to your lecturer.

|  |  |
| --- | --- |
| Student Signature |  |
| Date | 17/12/2023 |

1. **Data Understanding and Preparation**

This paragraph introduces the problem, the dataset and applies an exploratory data analysis (EDA) to understand the assignments’ objectives.

Sentiment Analysis (SA) is a sub field of text mining and is used to extract the opinions, sentiments and subjectivity of text [1]. SA uses lexical/statistical and deep learning methods to classify the sentiment of a given text [11]. The sentiments of a text give business owners valuable insights of the customer’s feedback. Due to the vast amount of product reviews, it requires high performance algorithms to extract the data of unstructed text reviews. This assignment aims to classify the sentiment of user reviews based on their text features.

Figure 1: Class distribution

As dataset Amazon reviews are given with their corresponding sentiment. The dataset contains 20,000 rows. Each row consists of a user review and it's corresponding sentiments. The labels are encoded as integers, 1 = positive and 0 = negative. The dataset has unbalanced classes, with 15,233 positive reviews and 4,767 negative reviews (Figure 1). The reviews contain spelling mistakes, for example "btother", "realustic" and "acording" in the following two texts:

1. 'This is a one of the best apps acording to a bunch of people and I agree it has bombs eggs pigs TNT king pigs and realustic stuff'

2. 'this. is fun an time consuming. works great on my kindle fire I really like this game so does my btother'

To avoid spelling mistakes the python library autocorrect is used. Autocorrect corrects the two texts as follows:

1. 'This is a one of the best apps according to a bunch of people and I agree it has bombs eggs pigs NT king pigs and realistic stuff'

2. 'this. is fun an time consuming. works great on my kindly fire I really like this game so does my brother'  
  
Even if the autocorrect library properly corrected all the mentioned mistakes, it wrongly corrected "TNT" to "NT" and "kindle fire" to "kindly fire" so this autocorrection is not used in this assignment. As second approach TextBlob is used:

1. "His is a one of the best apes according to a bunch of people and I agree it has bombs eggs pigs TNT king pigs and realistic stuff"

2. "this. is fun an time consuming. works great on my kindle fire I really like this game so does my brother"

TextBlob also makes mistakes by falsely correcting "apps" to "apes". Since auto-spelling correction is a challenging task and leads to misleading results as seen above, no spelling correction is applied for the review texts. Besides spelling mistakes the reviews contain letters that are repeated multiple times, like "soooooo". Tokens like "sooooo" likely occur just once in the corpus and the meaning of the token can hardly be extracted. Further, missing whitespaces between the punctuation and the new sentence makes the sentence recognition challenging, , e. g. "[...] make what this is all about.It's not relevant […]".

As an overview of the data quality and basic statistics of the dataset is given, the next step is to clean the the text data. Common practice to clean texts are punctuation and stopwords removal. However, [2] compare punctuation removal in sentiment analysis scenarios of english and turkish reviews and come to the conclusion to keep punctuation and stopwords because they contain meaningful information in terms of their sentiment. [[3]] also identifies punctuation as a good feature for Twitter sentiment analysis. Therefore, the stopwords and the punctuation are kept in the corpus, only lowercasing is applied in the cleaning process.

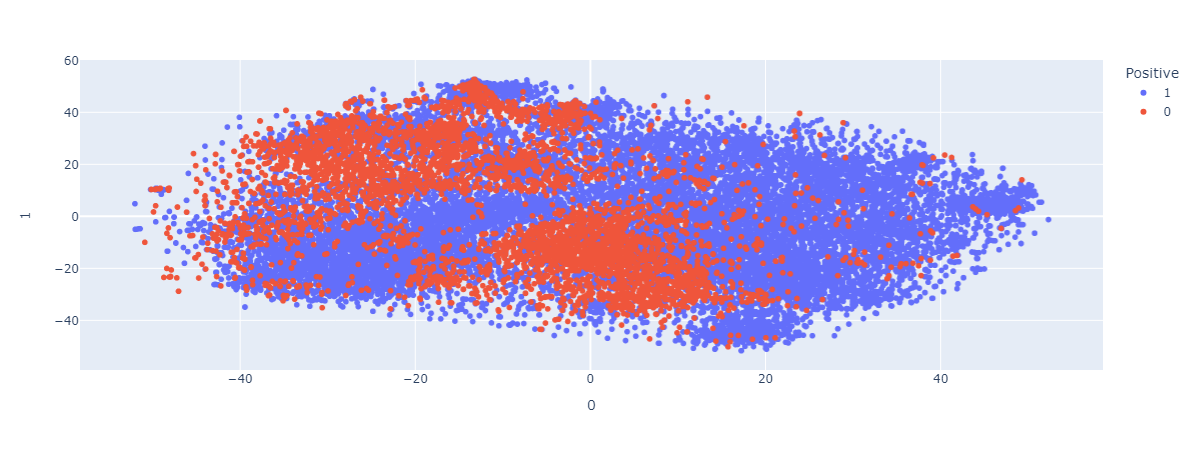
In [4], Transformer models outperform other machine learning models (i.e. LSTM, BoT, CNN) in sentiment analysis. Transformer is one of the most robust state-of-the-art approach in NLP. Therefore, the transformer model BERT is used to tokenize the reviews and to obtain BERT's 768-dimensional embeddings.

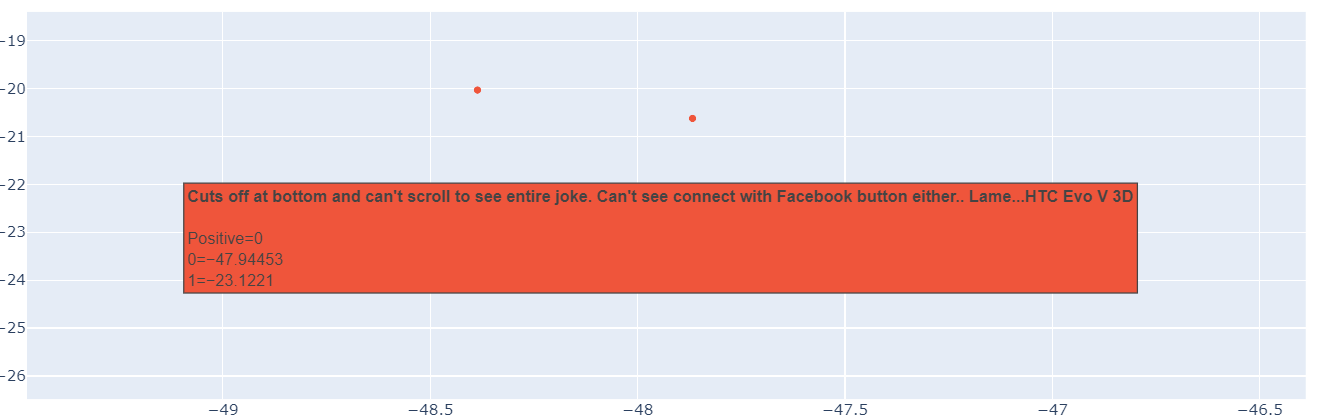
1. **Data Visualization and Exploration**

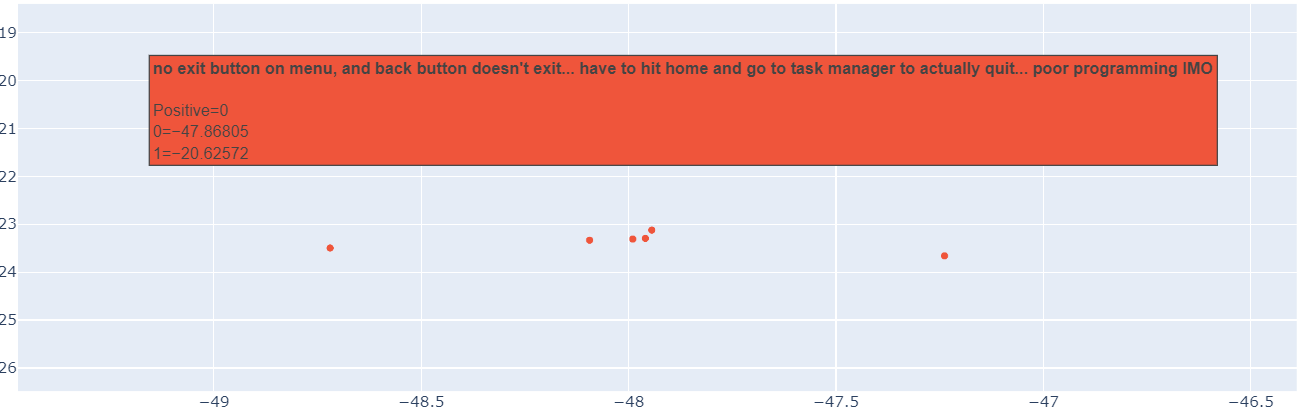
In this section, the computed 768-dimensional BERT embeddings are reduced to a 2-dimensional space using the t-distributed stochastic neighbour embedding (t-SNE). By conducting dimension reduction, we aim to find similar documents that are close to each other in a 2-dimensional vector space while preserving the meaning of the high dimensional embeddings [5]. We also could compare documents with the highest distances, e. g. the highest and lowest x-values to interpret the x dimension, but this data exploration focuses on similar documents that are represented in the vector space by similar x and y coordinates.

The result of all data points is a mixed bubble of positive and negative reviews (Figure 2) so we take a closer look at specific areas in the plot. In the range of (-49, -26) and (-47,-19) you can find a small cluster of negative labelled documents that are all about technical issues, like installation, crashes and technical features that doesn't work properly (Figure 3, 4, 5). In the vector space between (-52.1, -20) and (-51.6, 10) you find a positive document cluster whose reviews praise the alarm clock app (Figure 6, 7, 8). The points (1.13, 2.7) and (1.16, 2.75) don't share a consistent sentiment, but both reviews relate to a translation app (Figure 9, 10).

There are more document clusters that can be explored but that goes beyond the scope of this data exploration task.

Figure 2: t-SNE all data points

Figure 3: Negative cluster data point 1

Figure 4: Negative cluster data point 2

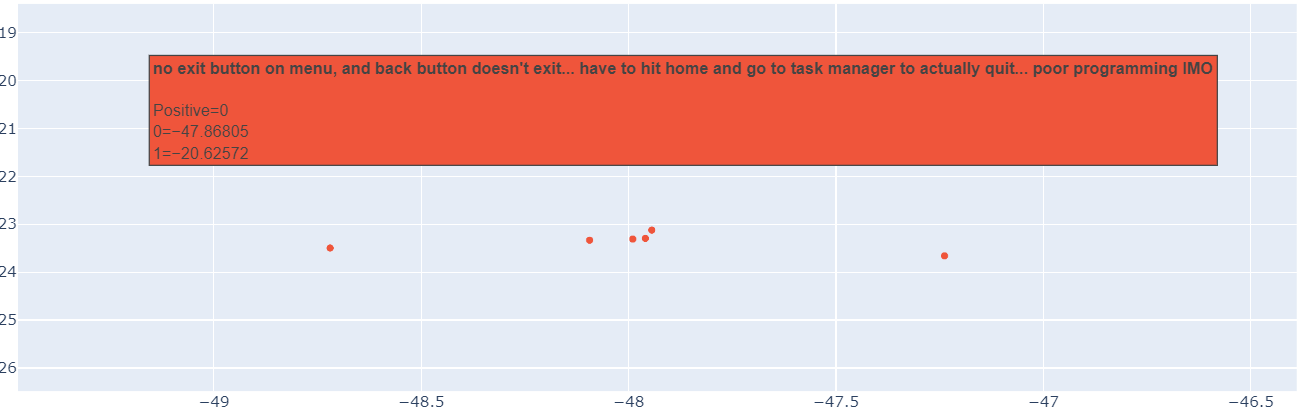
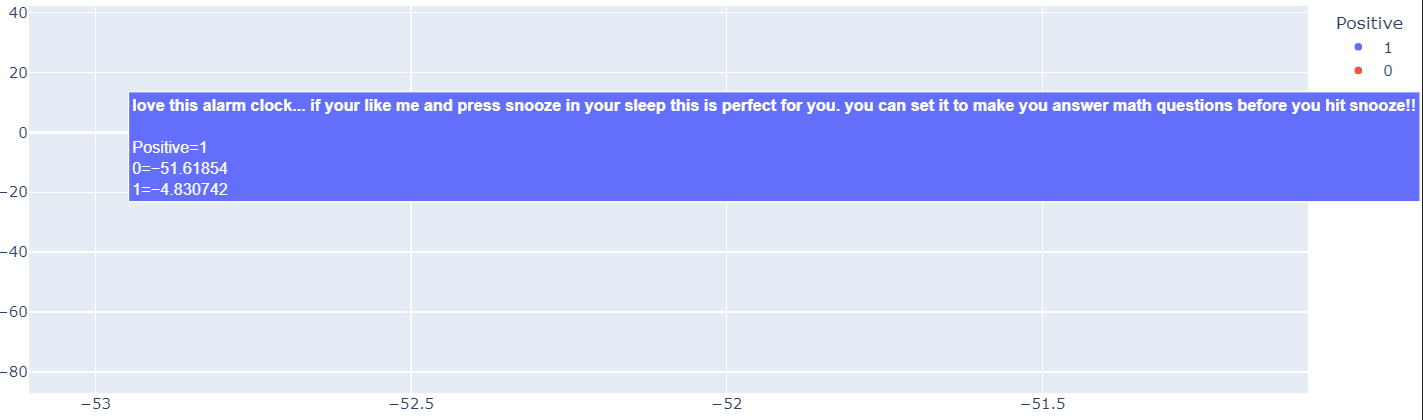
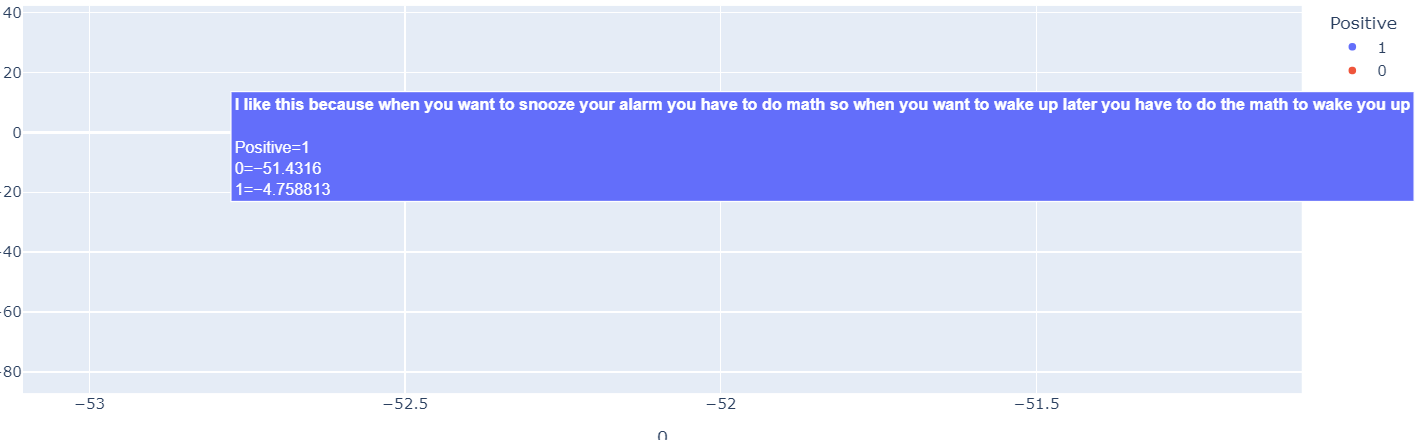
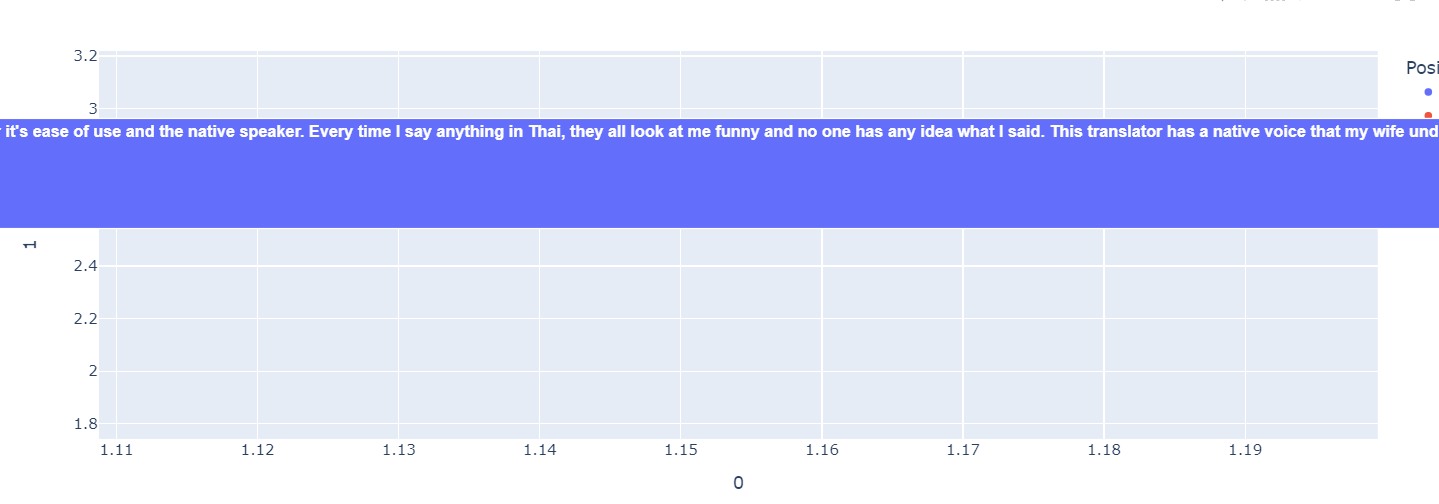
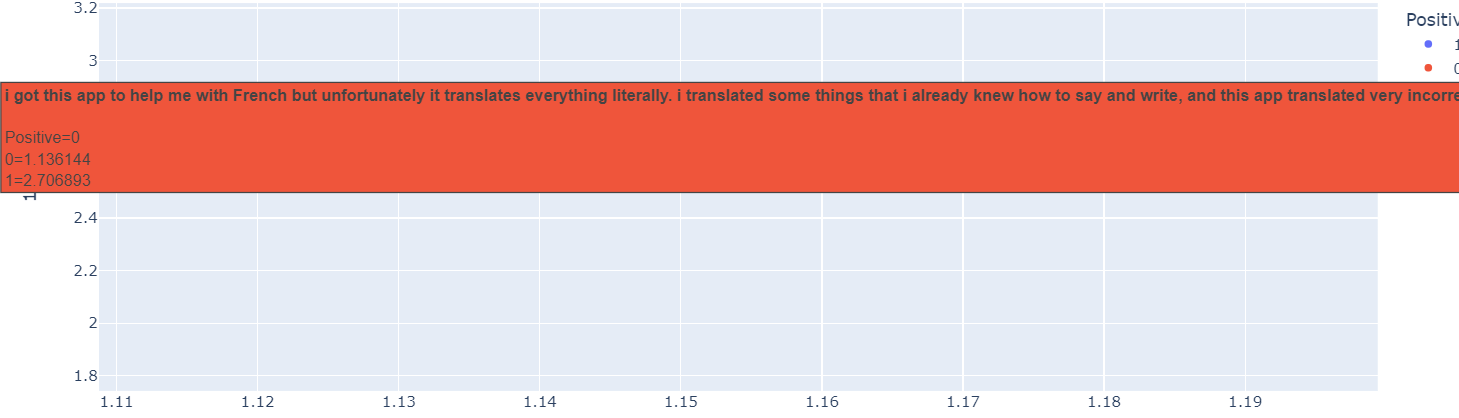
Figure 5: Negative cluster data point 3

Figure 6: Positive cluster data point 1

Figure 7: Positive cluster data point 2

Figure 8: Positive cluster data point 3

Figure 9: Positive review about translator app

Figure 10: Negative review about translator app

Next, we create two word clouds, one for the positive reviews and one for the negative reviews. The word clouds scale the tokens by their frequency, the larger the token the higher the frequency. As shown in Figure 11, you can see that the tokens "love" and "great" occur in positive reviews. "App", "play", "kindle fire" and "game" occurs in both positive and negative reviews (Figure 12) and are roughly equal in size. In the negative sentiment word cloud you can see tokens, among others "waste", "boring" and "stupid" which don't exist in the positive sentiment word cloud. One more interesting observation is that "fun" occurs in both word clouds but in the positive word cloud it is larger than in the negative word cloud.

Figure 11: Positive sentiment word cloud

Figure 12: Negative sentiment word cloud

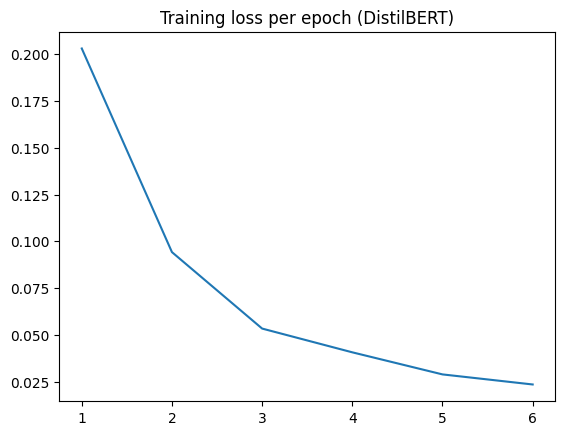
1. **Model Building and Evaluation**

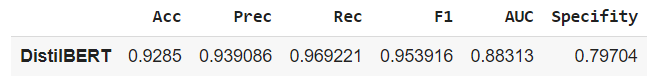
Building upon the tokenization and preprocessing steps, in this section a BERT model is fine-tuned on to the sentiment analysis task and evaluated. Therefore, the prepared training dataset is used to fine-tune a pretrained DistilBERT. DistilBERT is smaller and faster than a BERT model by using a compression technique of knowledge distillation [6].

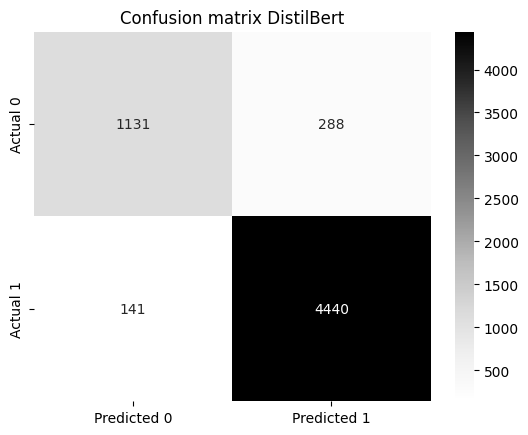
The following training parameters are applied for fine-tuning:

* Batch size: 8
* Optimizer: Adam
* Number epochs: 6
* Learning rate: 5e-5
* Loss function: BCELoss

This model was supposed to evaluated against SieBERT. SieBERT is specificially fined-tuned and evaluated on 15 different sentiment analysis datasets [7]. The text types ranges from tweets to various reviews. Since SieBERT is fine-tuned on binary sentiment analysis datasets, it requires less training data and less computation demands to attain an similar accuracy scores like a general-purpose language model. Unfortunately the computational resources for this assignment didn’t fulfill the requirements for training this model so only DistilBERT was fine-tuned on the training data.

Figure 13: Training loss per epoch

Figure 14: Evaluation metrics

Figure 15: Confusion matrix

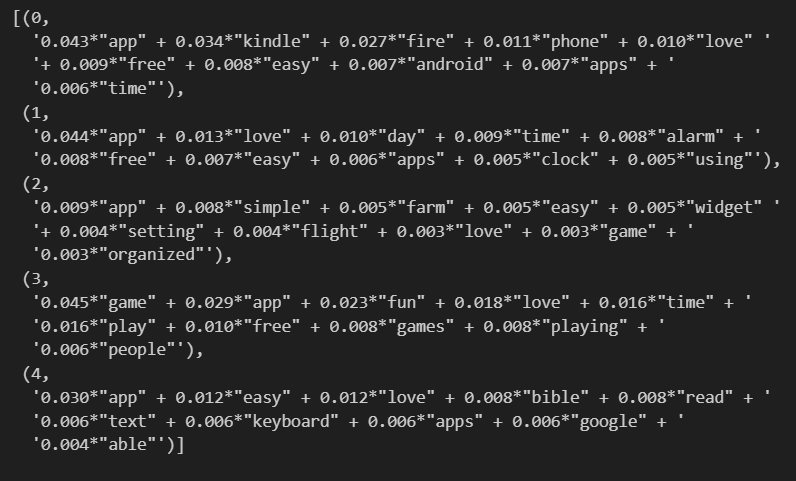
Regarding the training time, a DistilBERT model was fine-tuned in 5 min/epoch and evaluated in only 45 sec. DistilBERT achieves an overall accuracy of 92.85 %.The evaluation metrics and confusion matrix (Figure 14, 15) show positive reviews are better predicted (96.92 %) than negative reviews (79.7 %) which might be caused due to an imbalances of classes in the data set.

1. **Advanced Techniques and Analysis**

This section includes further text mining techniques: Latent Dirichlet Allocation (LDA) for topic modeling and k-Means for clustering reviews. Similar to section 3, the machine learning models are trained and evaluated to assess their quality and effectiveness.

LDA is a topic modeling algorithm that makes use of dimension reduction. Compared to T-SNE of section 2, LDA assumes that n latent topics are distributed randomly in each document where each topic is distinguished by a distribution of words [8].

For the LDA model, the reviews are tokenized, punctuation is removed and the stop words of the provided stop word list are removed. Additionally tokens containing an apostrophe like "'ve" or "\'s" are removed which have high impact on the results. The LDA model results in 5 topics, and the first 10 top keywords of each topic are shown in Figure 16.

Figure 16: LDA topics

"App" is represented in every topic with a high weight. You also see "apps" in topic 0, 1 and 4 because no stemming was applied in the preprocessing step. Keywords in topic 4 suggests to be a topic about reading and books, since it includes "bible", "read" and "text". Topic 1 seems to be about an alarm/clock app, which we've already discussed in section 2 by exploring the t-SNE. It is also difficult to extract meaningful topics from the top 10 keywords of each topic, especially because the topics have a large overlap. In further experiments the number of topics could be modified or the dictionary could be minimized by a specific POS-tag, e. g. only nouns are considered.

Clustering involves grouping similar data points together. K-means is a widely used clustering algorithm that partitions data into K clusters based on similarity. It is often applied in text analysis to discover patterns and group similar documents [9]. In this assignment, k-means is applied to cluster similar reviews by using BERT embeddings. To find an optimal k for the clustering algorithm, the elbow method is applied. The idea behind the elbow method is that one should choose a k that adding another cluster (k+1) doesn't give much better modelling of the data [10]. To inspect this, k is plotted against the variance explained by the clusters (Figure 17).

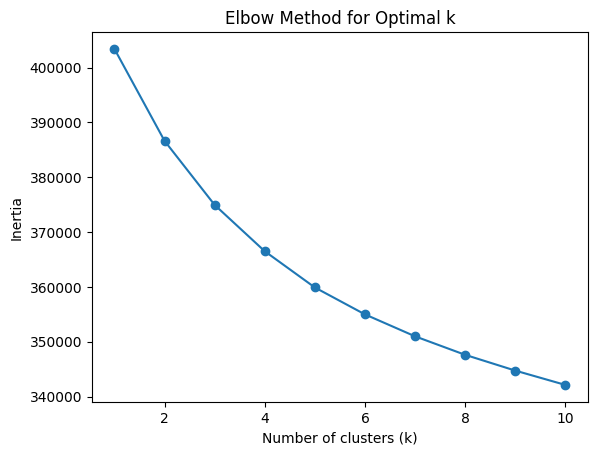
Figure 17: Elbow method

Figure 17 shows that there is no optimal value for k, as the variance explained constantly decreases. Therefore the clustering algorithm is utilized with k=2 to investigate whether k-Means can distinguish between positive and negative reviews.

|  |  |  |
| --- | --- | --- |
|  | Cluster 1 | Cluster 2 |
| Positive | 10,500 | 4,608 |
| Negative | 1,995 | 2,772 |
| Total | 12,620 | 7,380 |

Table 1: Cluster distribution

The resulting distribution of cluster and class labels are represented in Table 1. Both clusters don't seem to distinguish very well in terms of sentiment. Next, we'll take a closer look at the reviews that has the widest and the closest distances to each cluster centre. The following python output shows the widest and closest data points to each cluster with their corresponding euclidean distance at the end of each review:

Cluster 1 widest:

free app free app first commentfree free free free free free free free free free free free free apppppppppppp of the day 8.92

Uninstalled 7.82

Remove app from Digital Device 7.64

Cluster 1 closest:

I love this app! I like that I can write with my finger or type on the page if I choose. Since I type fast it would be great if the app wouldn't lag behind. I recommend this app for anyone who likes to keep notes or journals. I currently have this app ru 2.66

This App is so worth the price. Now I got mine on the Free App of the day however, I wouldn't mind paying for it. This App is endless...by that I mean, you can use your own pictures to create a puzzle. The App keeps track of your completion time and n 2.63

Beautiful and full size pictures/puzzles on my tablet (running Honeycomb 3.2). It is intuitive and easy to play. I didn't get this app for excitement, but rather because I enjoy puzzles. The developer didn't put in unnecessary permissions. To top it 2.59

Cluster 2 widest:

Fun and interesting to read. \_\_\_\_\_ \_\_\_\_ \_\_\_\_ \_\_\_\_ \_\_ \_\_ \_\_\_\_ \_\_\_\_\_ \_ \_\_\_ \_\_\_\_\_ \_\_\_\_ \_\_\_\_\_ \_\_\_\_ \_\_\_ 12.89

app is very usefull i got my amazon market debt paid off.. ;) good layout,...................................................................................................................................................................... good good app 11.2

dd d s s s s s ss s s s s s s s d s. s s s d s d dd d d d x xx 11.04

Cluster 2 closest:

I admit I&apos;m a free app whore. I download them then never review them. This app deserves a review though. It has amazing graphics and works great on Thunderbolt. Like Google Earth only a million times better!! I wish I was a great reviewer but alas I 2.98

look if you&apos;re going to download this app you better like old time radio stories. people voting one star because they think these kinds of stories are lame have no business downloading let alone reviewing this app. that said, this is a very easy to 2.96

I have &#34;Farkle Addict&#34; on my I phone and I LOVE that game. This one does not allow you to play a 1 person game and that is how I enjoy playing Farkle. Usually play it when I don't have much time to play, like when waiting in a Dr office. I wish y 2.89

Both clusters have as widest data points as widest data points, reviews that contain shorter texts and less information compared to the closest data points. The two closest data points to cluster 1 address a puzzle app. No patterns can be derived by the closest data points for second cluster centre.

1. **List of Insights, Recommendations and Future Works**

Section 1 outlines basic statistics and challenges regarding the data quality of amazon reviews. The dataset has unbalanced classes, with the majority belonging positive reviews and the minority class belonging to negative reviews. The deficiencies of the dataset like spelling mistakes and missing whitespaces have been analysed. Further, section 1 discusses cleaning processes and concludes that stopwords and punctation should not be removed, since those features contain meaningful information for a sentiment analysis task. Section 1 also describes the tokenization and selection of appropriate embeddings for sentiment analysis.

Section 2 includes the dimension reduction technique t-SNE and discusses its results. Positive and negative sentiment clusters could be derived that relate to common topics. More similarities between the reviews can be investigated and also the dimensions could be interpreted in future research. The word clouds show on a term frequency level which words are representative for each class. Among the positive reviews words like “love” and “great” occur frequently, while “bad”, “waste” and “boring” highlight important keywords in negative reviews.

In section 3, a pretrained BERT model is fine-tuned using the amazon reviews. The model achieves an accuracy of 92.85 % and predicts positive sentiments with a certainty of 96.92 %. The class imbalance probably causes weaker performance in classifying negative reviews (79.70%). To balance the imbalanced data set, resampling techniques such as oversampling or undersampling could be applied in the preprocessing step. Since DistilBERT is fine-tuned only in six epochs, a higher number of epochs could be selected in further experiments. Additionally this work strongly recommends the comparison to other transformer models like SieBERT as discussed in section 3, or to other architectures like GPT.

Section 4 discusses advanced techniques including LDA, a topic modeling method, and k-Means, an unsupervised clustering algorithm. As discussed in this section, no optimal k for k-Means could be derived, therefore only two clusters are computed. The data points with the smallest and largest distance to the cluster centre point are analysed. Further work can be carried out by experimenting with different k values or in finding differences between the two clusters. By applying LDA, 5 topics are extracted and the top 10 keywords are compared. As the topics are difficult to interpret due to overlapping keywords, the number of topics could be modified or the dictionary could be minimized by specific POS-tags in further experiments. Frequently occurring keywords such as “app” or “game” could also be removed from the dictionary before applying LDA in order to represent topics in more detail.

**References**

[1] Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, Volume 5, Issue 4, 2014, Pages 1093-1113, ISSN 2090-4479, <https://doi.org/10.1016/j.asej.2014.04.011>. (<https://www.sciencedirect.com/science/article/pii/S2090447914000550>)

[2] Parlar, Tuba & Özel, Selma & Song, Fei. (2019). Analysis of data pre-processing methods for the sentiment analysis of reviews. Computer Science. 20. 123. 10.7494/csci.2019.20.1.3097.

[3] Koto, F., Adriani, M. (2015). A Comparative Study on Twitter Sentiment Analysis: Which Features are Good?. In: Biemann, C., Handschuh, S., Freitas, A., Meziane, F., Métais, E. (eds) Natural Language Processing and Information Systems. NLDB 2015. Lecture Notes in Computer Science(), vol 9103. Springer, Cham. <https://doi.org/10.1007/978-3-319-19581-0_46>

[4] Balci, S., Demirci, G.M., Demirhan, H., Sarp, S. (2022). Sentiment Analysis Using State of the Art Machine Learning Techniques. In: Biele, C., Kacprzyk, J., Kopeć, W., Owsiński, J.W., Romanowski, A., Sikorski, M. (eds) Digital Interaction and Machine Intelligence. MIDI 2021. Lecture Notes in Networks and Systems, vol 440. Springer, Cham. <https://doi.org/10.1007/978-3-031-11432-8_3>

[5] Huroyan, Vahan & Navarrete, Raymundo & Hossain, Md & Kobourov, Stephen. (2022). Embedding Neighborhoods Simultaneously t-SNE (ENS-t-SNE). 10.48550/arXiv.2205.11720.

[6] Chan, J.YL., Bea, K.T., Leow, S.M.H. et al. State of the art: a review of sentiment analysis based on sequential transfer learning. Artif Intell Rev 56, 749–780 (2023). https://doi.org/10.1007/s10462-022-10183-8

[7] Jochen Hartmann, Mark Heitmann, Christian Siebert, Christina Schamp, More than a Feeling: Accuracy and Application of Sentiment Analysis, International Journal of Research in Marketing, Volume 40, Issue 1, 2023, Pages 75-87, ISSN 0167-8116, <https://doi.org/10.1016/j.ijresmar.2022.05.005>. (https://www.sciencedirect.com/science/article/pii/S0167811622000477)

[8] Mohammed, Shaymaa & Al-augby, Salam. (2020). LSA & LDA Topic Modeling Classification: Comparison study on E-books. 2502-4752. 10.11591/ijeecs.v19.i1.pp%25p. \

[9] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.

[10] Bholowalia, Purnima, and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN." International Journal of Computer Applications 105.9 (2014).

[11] Lopez, G., Nguyen, A., & Kaul, J. (2023). Reducing Computational Costs in Sentiment Analysis: Tensorized Recurrent Networks vs. Recurrent Networks. ArXiv, abs/2306.09705.