# Text Mining and Sentiment Analysis
# How Do You Feel, My Dear

Fatma Er, 967585

June 9, 2023

**Abstract**

The goal of this project is to develop a model that can predict emotions in text using the Emotion Detection from Text dataset as a training set. The emotions can be represented as categorical classes or in a continuous space such as ValenceArousal-Dominance. The model will be used to analyze the emotional profiles of main characters in a movie from the Cornell Movie-Dialogs Corpus, and to observe how these profiles change over time and are influenced by the characters' relationships.

**Key words:** Emotion detection, Sentiment analysis, Fictional scripts, Movies, TV series, Cornell Movie-Dialogs Corpus, Character analysis, Relationship analysis, Natural language processing.

## 1 INTRODUCTION

Emotion detection in text has become a topic of significant interest in natural language processing research in recent years. The ability to accurately identify and understand the emotions expressed in text data has numerous applications in different domains, including marketing, social media analysis, customer feedback analysis, and mental health diagnosis. By analyzing emotions in text, researchers can gain valuable insights into human communication and behavior, as well as track changes in sentiment over time. To achieve this goal, text mining and sentiment analysis techniques have become increasingly important for understanding the complex nature of emotions in text.

In this paper, we present a model for predicting emotions in text using machine learning and deep learning algorithms. Specifically, we use the Emotion Detection from Text dataset as a training set to develop our model. The emotions can be represented as categorical classes or in a continuous space such as Valence-Arousal-Dominance. We evaluate the performance of our model using accuracy score.

To demonstrate the applicability of our model, we use it to analyze the emotional profiles of main characters in a movie from the Cornell Movie-Dialogs Corpus. We observe how these emotional profiles change over time and are influenced by the characters' relationships. Our results show that our model can accurately predict emotions in text, and that the emotional profiles of the characters in the movie can provide valuable insights into their behavior and relationships.

Overall, our paper contributes to the growing body of literature on emotion detection in text, and demonstrates the potential applications of this technology in different domains. By accurately predicting emotions in text, we can gain a deeper understanding of human communication and behavior, and improve the effectiveness of various applications.

In this project, we have developed several deep learning algorithms such as Random Forest , Long-Short Term Memory (LSTM), Gradient Boosting and Logistic Regression algorithms were used for estimation. The input to our models are Twitter Dataset. We trained the models with the twitter's emotion data and output the predicted category label that a specific emotion belonged to. We compared the performance of multiple algorithms with different features and explored and predict ConvoKit Dataset movie diyologs and character's emotion during movie.

The paper is organized as follows: Section II describes the related work and Section III introduces the dataset and features we are using in this project. The methods, models and model's average accuracy are presented in section IV, followed by the section V presenting Conclusion about our paper.

# 2 RELATED WORKS

In this section, the studies published in the literature Emotion Detection from Text are examined.

(Binali et al., 2010) explores the use of computational approaches for emotion detection in text. The authors review existing research on emotion detection, including feature-based approaches, machine learning algorithms, and deep learning methods, and discuss the importance of context in emotion detection. The article proposes a hybrid architecture for emotion detection using a support vector machine (SVM) algorithm and achieves a high prediction accuracy of 96.43% on web blog data. The authors also discuss potential applications of emotion detection technology in customer feedback analysis, mental health diagnosis, and social media monitoring. Overall, the article highlights the potential of computational approaches for detecting emotions in text and their growing importance in various domains.

Emotional valence and arousal have a significant impact on word recognition, yet current models of word recognition often overlook these factors. The study by (Kuperman et al., Feb 2014) aimed to determine the precise nature of the effects of valence and arousal on word recognition using a large sample of words and controlling for lexical and semantic factors. The results revealed that valence and arousal exert independent monotonic effects on word recognition, with negative and arousing words recognized more slowly than positive and calming words. Valence and arousal do not interact, but both interact with word frequency. Incorporating emotional factors, particularly valence, can improve the performance of models of word recognition. These findings highlight the importance of considering emotional factors in understanding the cognitive mechanisms underlying emotional word recognition.

# 3 DATASET AND FEATURES

## 3.1 Emotion Detection from Text Using a Twitter Dataset

The dataset provided is a valuable resource for addressing the challenging problem of emotion detection from text in Natural Language Processing. With 40,000 tweets annotated with their corresponding emotions, we have a labeled dataset that can be used to train and test machine learning models for this task.

The dataset contains 13 different emotions, including neutral (8638), worry (8459), happiness (5209), sadness (5165), love (3842), surprise (2187), fun (1776), relief (1526), hate (1323), empty (827), enthusiasm (759), boredom (179), and anger (110). These emotions reflect the diverse range of human experiences and can provide valuable insights into the sentiments expressed in text data. By accurately detecting these emotions, we can better understand the attitudes and opinions of individuals and groups, and use this information for a variety of applications, such as marketing, customer feedback analysis, and social media monitoring.

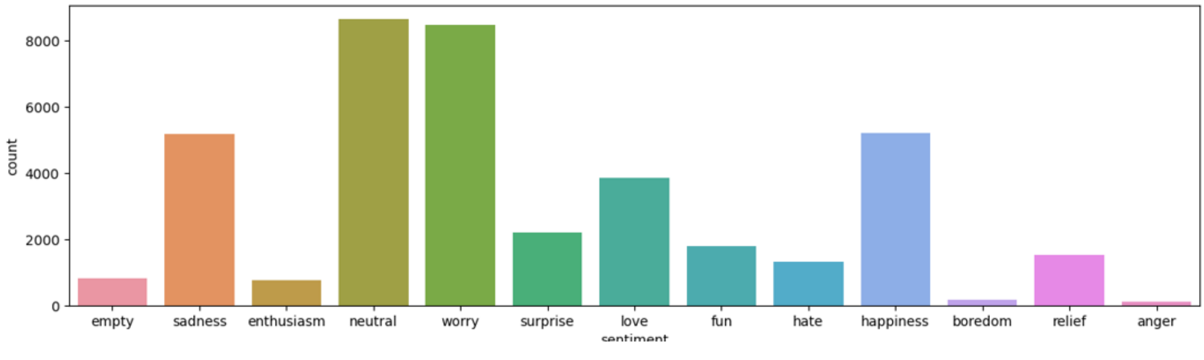


Figure 1: Emotion Distribution

Figure 1 showed us Emotion Distribution for tweets.

Balancing Data Because emotion data is so far apart, for example, neutral contains 8638 data, whereas anger contains 110 data. We then decided to balance the data to make the data fairer in terms of learning accuracy. We use the Oversampling method, which means adding synthetic data, which refers to the largest amount of data in the dataset.

Thus, all emotions reached an equal number(8638).

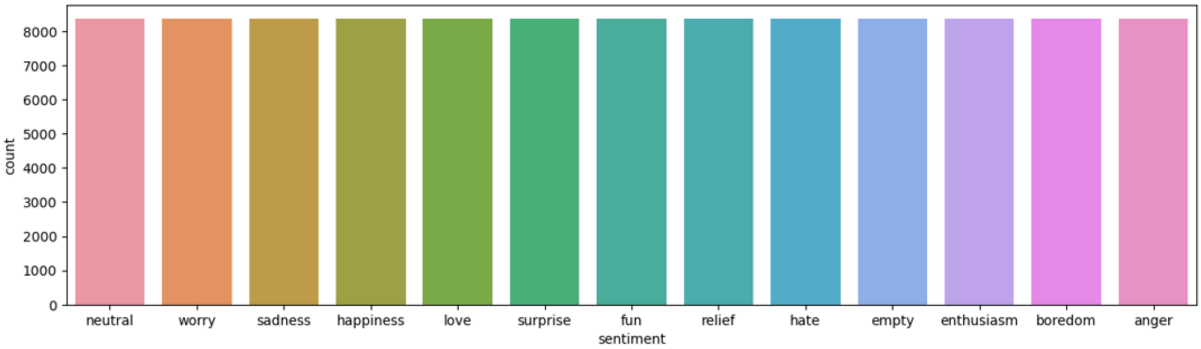


Figure 2: Emotion Distribution Augmented

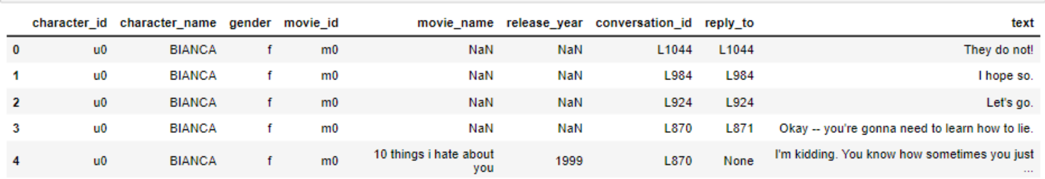
## 3.2 The ConvoKit Dataset

ConvoKit is an open-source toolkit that facilitates the analysis of conversations across various domains, such as social media, online forums, messaging apps, face-to-face institutional interactions, and even fictional exchanges. It includes pre-formatted datasets that cover a wide range of conversational settings, such as Supreme Court transcripts, Wikipedia talk pages, Reddit discussions, and movie dialogs. These datasets demonstrate the versatility of ConvoKit's conversation representation.

Moreover, ConvoKit provides guidelines and code for transforming other datasets into ConvoKit format, which expands the toolkit's capabilities to analyze conversational data beyond the pre-formatted datasets. The toolkit offers several features, including advanced natural language processing capabilities, support for multiple data formats, and a suite of visualization tools. It is designed to be user-friendly, with extensive documentation and a user-friendly interface, and has a large and active community of users and developers who contribute to its ongoing development and maintenance.

The Cornell Movie-Dialogs Corpus is one of the pre-formatted datasets available in ConvoKit. It consists of conversations extracted from movie scripts, and it includes over 220,579 conversational exchanges between 10,292 character pairs from 617 movies. The dataset covers a variety of genres, including action, comedy, drama, and sci-fi.

The Cornell Movie-Dialogs Corpus serves as a valuable resource for natural language processing and dialogue research. It allows for the analysis of conversational dynamics, such as turn-taking and topic shifts, and the development of machine learning models for dialogue generation and response prediction.



| | character_id | character_name | gender | movie_id | movie_name | release_year | conversation_id | reply_to | text |
|---|---|---|---|---|---|---|---|---|---|
| 0 | u0 | BIANCA | f | m0 | NaN | NaN | L1044 | L1044 | They do not! |
| 1 | u0 | BIANCA | f | m0 | NaN | NaN | L984 | L984 | I hope so. |
| 2 | u0 | BIANCA | f | m0 | NaN | NaN | L924 | L924 | Let's go. |
| 3 | u0 | BIANCA | f | m0 | NaN | NaN | L870 | L871 | Okay -- you're gonna need to learn how to lie. |
| 4 | u0 | BIANCA | f | m0 | 10 things i hate about you | 1999 | L870 | None | I'm kidding. You know how sometimes you just ... |

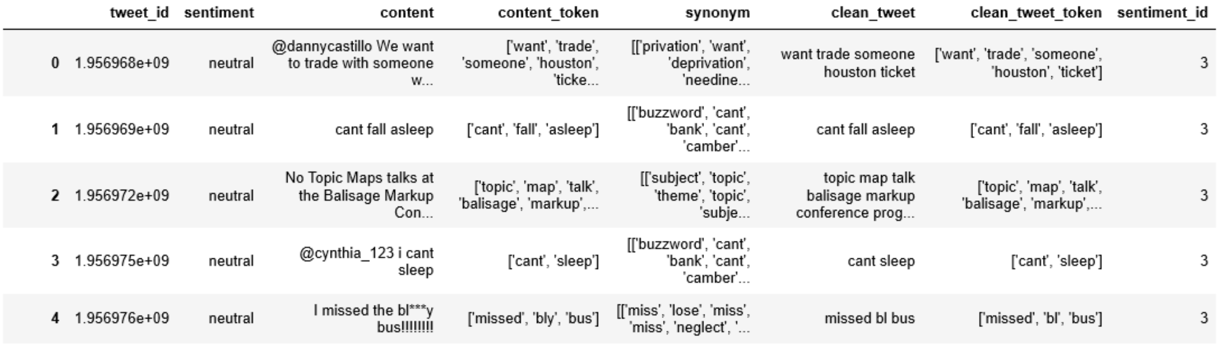
Figure 3: Convokit Data Example

## 3.3 Data Processing and Cleaning Text

In order to prepare the text data for analysis, we need to apply natural language processing (NLP) techniques.The NLP steps we applied to the data are Tokenization, Normalization, Lemmatization, and Stopword Removal.

Tokenization refers to the process of splitting the text into individual words or tokens. Normalization is used to convert all the text to lowercase to ensure consistency. Lemmatization is the process of reducing words to their base form or lemma. Finally, stopword removal involves removing common words such as "the," "and," and "in" as they do not provide much meaning to the text.

By applying these NLP steps to the data, we can preprocess the text and transform it into a format that is suitable for further analysis.



| | tweet_id | sentiment | content | content_token | synonym | clean_tweet | clean_tweet_token | sentiment_id |
|---|---|---|---|---|---|---|---|---|
| 0 | 1.956968e+09 | neutral | @dannycastillo We want to trade with someone w... | ['want', 'trade', 'someone', 'houston', 'ticke... | [['privation', 'want', 'deprivation', 'needine... | want trade someone houston ticket | ['want', 'trade', 'someone', 'houston', 'ticket'] | 3 |
| 1 | 1.956969e+09 | neutral | cant fall asleep | ['cant', 'fall', 'asleep'] | [['buzzword', 'cant', 'bank', 'cant', 'camber'... | cant fall asleep | ['cant', 'fall', 'asleep'] | 3 |
| 2 | 1.956972e+09 | neutral | No Topic Maps talks at the Balisage Markup Con... | ['topic', 'map', 'talk', 'balisage', 'markup',... | [['subject', 'topic', 'theme', 'topic', 'subje... | topic map talk balisage markup conference prog... | ['topic', 'map', 'talk', 'balisage', 'markup',... | 3 |
| 3 | 1.956975e+09 | neutral | @cynthia_123 i cant sleep | ['cant', 'sleep'] | [['buzzword', 'cant', 'bank', 'cant', 'camber'... | cant sleep | ['cant', 'sleep'] | 3 |
| 4 | 1.956976e+09 | neutral | I missed the bl***y bus!!!!!!!! | ['missed', 'bly', 'bus'] | [['miss', 'lose', 'miss', 'miss', 'neglect', '... | missed bl bus | ['missed', 'bl', 'bus'] | 3 |

Figure 4: Tokenization text

Figure 4 displays the tokens obtained after performing NLP steps such as Tokenization, Normalization, Lemmatization, and cleaning stopwords. The resulting tokens are used for training and testing our model.

The word cloud that appeared in Figure 5 formed the love data to be train and test by our model.

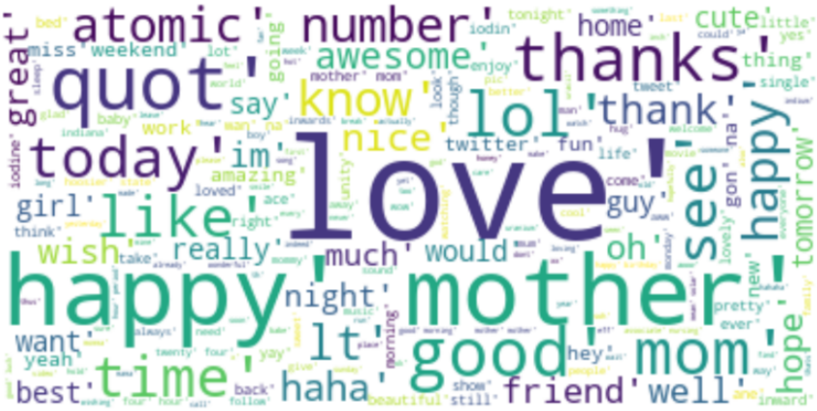
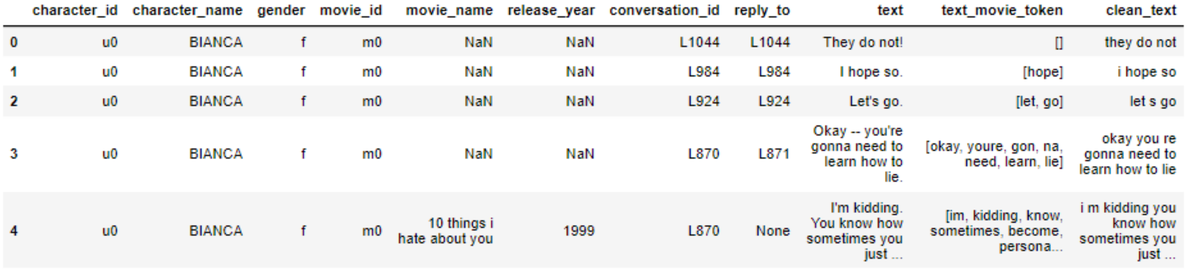


Figure 5: Love Emotion Word Cloud



| | character_id | character_name | gender | movie_id | movie_name | release_year | conversation_id | reply_to | text | text_movie_token | clean_text |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | u0 | BIANCA | f | m0 | NaN | NaN | L1044 | L1044 | They do not! | [] | they do not |
| 1 | u0 | BIANCA | f | m0 | NaN | NaN | L984 | L984 | I hope so. | [hope] | i hope so |
| 2 | u0 | BIANCA | f | m0 | NaN | NaN | L924 | L924 | Let's go. | [let, go] | let s go |
| 3 | u0 | BIANCA | f | m0 | NaN | NaN | L870 | L871 | Okay -- you're gonna need to learn how to lie. | [okay, youre, gon, na, need, learn, lie] | okay you re gonna need to learn how to lie |
| 4 | u0 | BIANCA | f | m0 | 10 things i hate about you | 1999 | L870 | None | I'm kidding. You know how sometimes you just ... | [im, kidding, know, sometimes, become, persona... | i m kidding you know how sometimes you just ... |

Figure 6: Conkovit Data for Prediction

The text that appeared in Figure 6 formed the data to be predicted by our model.

# 4 MODELS, RESULTS AND DISCUSSION

At this stage, the best model is defined by comparing the results obtained from the developed four models.

## 4.1 Comparison of The Results

In our study, ConvoKit dataset containing 304446 film conversation text . Random Forest , Long-Short Term Memory (LSTM), Gradient Boosting and Logistic Regression models were applied to the collected data. The overall results of these models are given in Table 1. It is clearly seen that Random forest is the best model value was obtained extraction with a success rate of 68 %

Table 1: Comparison of the results obtained from developed models

| Model | Accuracy (%) |
|---|---|
| Random Forest | 68% |
| LSTM | 68% |
| Logistic Regression | 61% |

## 4.2 Movie Emotion Prediction with Random Forest Model

With movies character emotion with rendom forest we choosed simple two person emotion during the movie. Our aim is here how is good our model for prediction the people feeling during conversation.

movie :10 things i hate about you and speaker: BIANCA conversation feeling distribution included 98 line text.

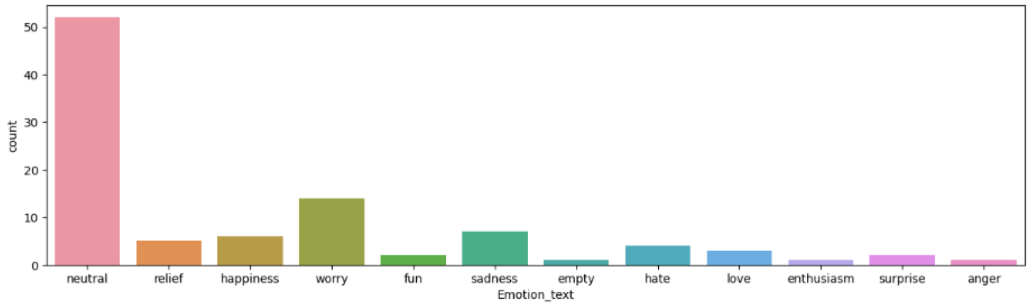


Figure 7: movie :10 things i hate about you and speaker: Bianca

movie :Clerks and speaker: DANTE conversation feeling distribution included 537 line text.

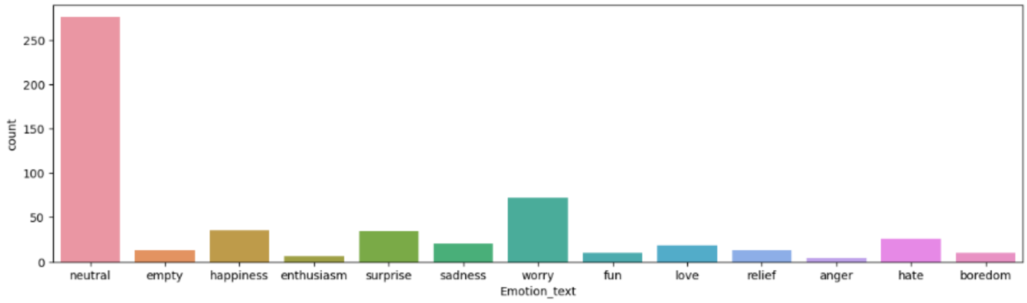


Figure 8: movie :Clerks and speaker: Dante

### 4.3 Discussion

As given in Table 1, our model offers result .The main reason for this part can be explained as follows;

1.The original dataset is not large/diverse enough for the models to perform well.

2.Usually, the separation of train and test data is done randomly.In our study, we can try different model seperation and use different model estimators. In future study we can go more deep. 3. In the future we can try also Bi-LSTM Networks model. It can be show us different model.

## 5 CONCLUSION

In this research paper, a movie emotion prediction system based on four algorithms has been proposed, and firstly, Twitter and ConvoKit data text study was carried out with text mining algorithms. Random Forest , Long-Short Term Memory (LSTM), Gradient Boosting and Logistic Regression algorithms were used for estimation. The average accuracy achieved in the test folds was reported and interpreted. Random Forest gave the best result , and the two characters selected with this model examined the change in emotion throughout the movie.

## 6 DECLARATION

I declare that this material, which we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

## References

H. Binali, C. Wu, and V. Potdar. Computational approaches for emotion detection in text. *Digital Ecosystems Business Intelligence Institute Curtin University of Technology Perth , Australia*, pages 1–7, 2010.

V. Kuperman, Z. Estes, M. Brysbaert, and A. B. Warriner. Emotion and language:

Valence and arousal influence word recognition. *Journal of Experimental Psychology General*, Feb 2014.