Fragrant Words: Analyzing the Intersection of Smell, Emotion, and Culture in the Middle East Through English Texts

submitted in partial fulfillment for the degree of master of science

Ali Alghamdi
14988828

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 19.04.2024

	UvA Supervisor	**External Supervisor**
Title, Name	Dr. Ali Alsahag	Dr. Marieke van Erp
Affiliation	UvA Supervisor	External Supervisor
Email	a.m.m.alsahag@uva.nl	marieke.van.erp@dh.huc.knaw.nl

ABSTRACT

The sense of smell is crucial in shaping human sensory experiences. Artificial Intelligence (AI) has seen significant advancements in identifying and understanding scents, but the focus has largely been on their chemical composition rather than exploring their textual descriptions. Exploring heritage data through semantic web, text mining, and image recognition offers a novel approach to uncovering olfactory experiences. The Middle East, with its deep cultural and historical significance, presents a rich area for such literary investigation. However, European narratives about Middle Eastern scents are probably affected by colonial biases. Additionally, there is a notable scarcity of studies on Middle Eastern olfactory experiences within the field. Therefore, this study sets out to analyze European olfactory heritage data related to the Middle East, focusing on the specific scents mentioned and their associated emotions utilizing sentiment analysis techniques. Utilizing Odeuropa's database, this research zeroes in on English-language texts about Middle Eastern olfactory experiences. The goal is to illuminate how European literature has depicted the Middle Eastern olfactory landscape, thereby contributing new insights into sensory heritage and deepening the understanding of how smell is interwoven with cultural and historical narratives.

KEYWORDS

odeuropa, olfaction, olfactory experience, nlp, plutchik, emotion detection, sentiment analysis, middle east

GITHUB REPOSITORY

https://github.com/alialghamdi/msc-thesis

1 METHODOLOGY

This methodology portion will provide an in-depth explanation of the research process, from the initial stages of data acquisition, cleansing, and formatting. Specifically, it will describe the collection and modification of Odeuropa's dataset to align with the required shape for analysis techniques. Furthermore, it will detail the execution data analysis on the dataset to uncover patterns and identify any limitations. The section will then elaborate on the process of selecting an appropriate model, focusing on how the dataset will be trained, and subsequently, how the model's performance will be assessed.

1.1 Dataset

1.1.1 Odeuropa Data Composition. First of all, the research plans to utilize the data from the Odeuropa dataset. The dataset is built on on CIDOC Conceptual Reference Model and that is event based, and in this case the smell is the event. The design of the data model of Odeuropa is in Figure 1. The dataset is currently hosted on a Knowledge Graph Database[1].

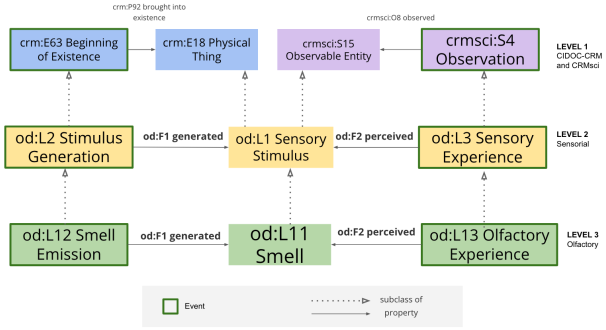

Figure 1: Odeuropa Data Model

1.1.2 Data Extraction. The author utilizes the exploration tool[2] provided by Odeuropa to efficiently navigate the dataset, specifically focusing on data pertaining to Middle Eastern countries, which aligns with the research area of this study. This tool features an export function that enables the extraction of data in both comma-separated values (CSV) and JavaScript Object Notation (JSON) formats. Due to the hierarchical structure of the dataset, which is inadequately represented in CSV format, JSON was chosen for its capability to preserve the nested data architecture effectively.

Consequently, the dataset subset extracted for this study encapsulates information relevant to Middle Eastern countries and is derived solely from the Odeuropa dataset. The linguistic composition of the subset mirrors the diversity of the original dataset. As part of the subsequent data cleaning process, the dataset will be filtered to include only English-language entries, as planned for this research.

1.1.3 Data Cleaning and Preprocessing. The data includes the excerpts of each text that reveal a specific smell, each row that is in the dataset, has a text related to it that is highlighted by the Odeuropa dataset as a description of a smell. As we import the exported dataset that is in JSON format, we start to eliminate all rows that are not in English language.

Upon importing the dataset in JSON format, the initial step in our data cleaning process involves isolating entries in English, thereby ensuring that the analysis remains focused on the primary language of the study. This filtering is crucial as it maintains a consistent linguistic base for subsequent textual manipulations.

Once we have a subset of English-only entries, specific cleaning techniques are applied exclusively to the text excerpts describing olfactory experiences. This selective approach is vital for maintaining the integrity and relevance of the data related to scent descriptions. Initially, all these excerpts are converted to lowercase to eliminate discrepancies caused by varied capitalization, which could affect later analyses. Punctuation within these excerpts is also removed,

[1] https://data.odeuropa.eu

[2] https://explorer.odeuropa.eu

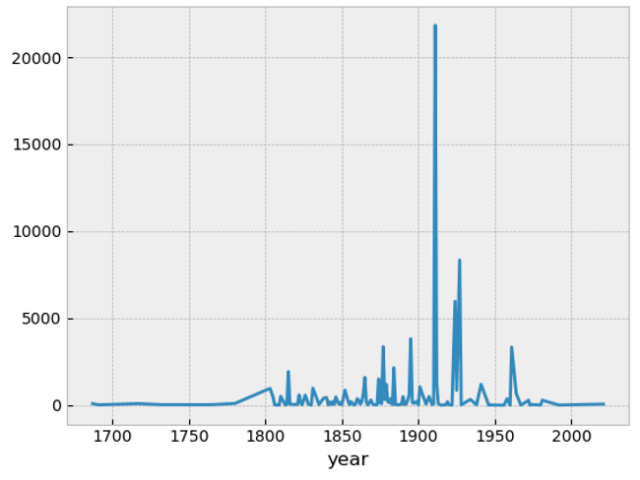

Figure 2: Distribution of the Years

simplifying the text structure and ensuring that the analysis can focus on the words themselves without interference from extraneous characters.

In addition to these modifications, common stop words are removed from these olfactory descriptions. This step is crucial as it helps in distilling the text to its most meaningful components, emphasizing words that directly contribute to the semantic richness of the scent descriptions. Furthermore, words are lemmatized to their root forms, ensuring that different variants of the same word are analyzed as a single entity, thereby simplifying the linguistic analysis and enhancing its accuracy. Finally, to aid in any temporal analysis, the dates associated with each text excerpt are checked and standardized, ensuring all entries follow a uniform date format. These meticulous steps in cleaning the data ensure that our analysis is based on precise and relevant textual and chronological information, paving the way for robust analytical outcomes.

1.1.4 Data Analysis. The dataset utilized in this research contains a total of 79,055 entries. These entries include 61,615 unique text excerpts that describe a diverse array of 332 distinct smells. Additionally, the dataset is drawn from 304 unique sources. These sources are varied and include books, historical artifacts, and other related materials. It is important to note that many of these entries repeatedly come from the same sources, indicating a concentration of olfactory descriptions in certain texts or artifacts.

The temporal coverage of the dataset spans from the year 1687 to 2021. This extensive period allows for a detailed examination of how descriptions of smells have changed over time, providing a valuable perspective on generational shifts in the perception of odors. The majority of the data, however, is clustered between the years 1800 and 19802. This concentration offers a rich set of data for studying olfactory references during the modern era.

To further understand the emotional context of these olfactory descriptions, a sentiment analysis was conducted using TextBlob, a straightforward lexicon-based tool. This analysis helped to identify and measure the emotional tones associated with different smells throughout the dataset. By applying this method, we observed

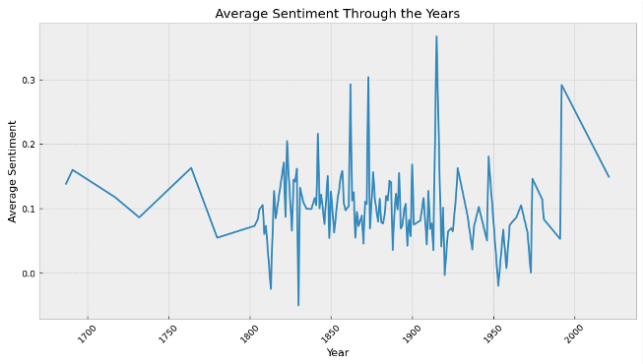

Figure 3: Average Sentiment Through the Years

how the sentiments expressed in these descriptions varied across different time periods3. This approach sheds light on the changing emotional landscapes associated with various odors as captured in the dataset over several centuries.

1.2 Sentiment Analysis

1.2.1 Relevant Models. When considering sentiment analysis models for olfactory experiences, several approaches stand out as potentially effective. Lexicon-based models, such as VADER and TextBlob, rely on predefined sentiment lexicons to assign sentiment scores to words and phrases. These models are straightforward to implement and can provide a quick overview of sentiment trends, but may struggle with context-dependent sentiments and complex linguistic structures, especially when dealing with the nuanced emotions represented in Plutchik's4 wheel of emotions[1]. Additionally, these models may not capture the subtle differences between the eight primary emotions and their varying intensities as described by Plutchik.

Alternatively, deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown promise in sentiment analysis tasks, including the detection of Plutchik's emotions. CNNs excel at extracting local features from text, such as key phrases indicating specific emotions, while RNNs are adept at capturing long-range dependencies and contextual information. These models can potentially learn the intricate relationships between words and phrases that contribute to the expression of different emotions in Plutchik's model. However, these models often require large annotated datasets for training, which may be a limitation for niche domains like olfactory experiences. Moreover, the complexity of Plutchik's emotion model, with its eight primary emotions and their combinations, may pose challenges in creating comprehensive annotated datasets.

Transformers, particularly BERT (Bidirectional Encoder Representations from Transformers), have revolutionized NLP tasks, including sentiment analysis and emotion detection based on Plutchik's model. BERT's ability to understand the bidirectional context of words in sentences makes it well-suited for capturing nuanced sentiments and identifying the specific emotions from Plutchik's wheel. Its pre-training on large, diverse datasets allows for effective fine-tuning on domain-specific tasks, even with limited annotated data.

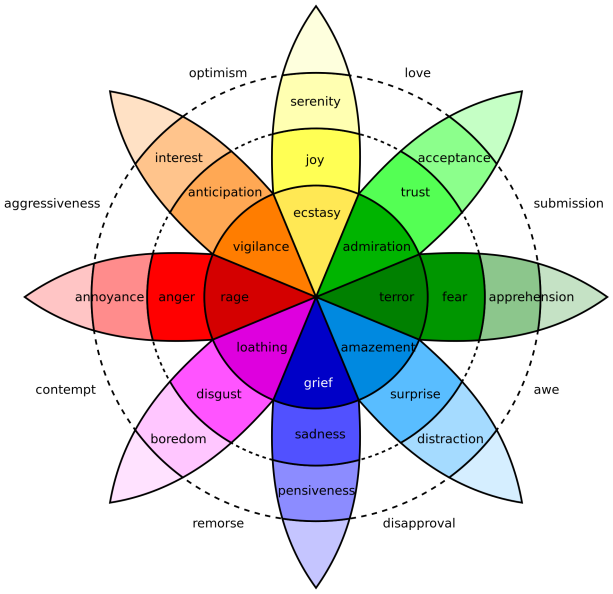

Figure 4: Plutchik's wheel of emotions

This is particularly advantageous when dealing with the complexity of Plutchik's emotion model, as BERT can learn to recognize the subtle linguistic patterns associated with each emotion and their combinations. Furthermore, BERT's attention mechanism can help identify the most relevant words and phrases contributing to the expression of specific emotions, enabling a more accurate analysis of the emotional content in olfactory descriptions.

1.2.2 Model Selection. In selecting an appropriate model for conducting sentiment analysis on olfactory experiences using Plutchik's emotions, we have chosen BERT due to its advanced capabilities in understanding the context of words in sentences. The decision to use BERT over other models is based on its state-of-the-art performance on a variety of NLP tasks, including sentiment analysis and emotion detection. This approach allows us to leverage BERT's understanding of language context and semantics, which is essential for accurately interpreting the sentiments and identifying the specific emotions from Plutchik's wheel associated with olfactory descriptions. BERT's ability to capture the nuances of language and its robustness in handling complex linguistic structures make it an ideal choice for analyzing the rich emotional content in the Odeuropa dataset. Moreover, BERT's pre-training on diverse datasets can help mitigate the limitations posed by the lack of large annotated datasets specific to olfactory experiences and Plutchik's emotions.

1.2.3 Model Training. For training sentiment analysis models to detect Plutchik's emotions on our dataset of olfactory experiences from Odeuropa, we will initiate the process by pre-processing the data to ensure it meets the input requirements of the chosen sentiment analysis techniques. This process includes tokenizing the text into a format comprehensible by the models, which involves breaking down the text into tokens and associating them with their respective indices in the model's vocabulary. Additionally, we will explore techniques such as data augmentation to expand our training dataset and improve the model's ability to generalize across diverse olfactory descriptions. This may involve generating synthetic examples by applying semantic transformations or using domain-specific synonyms to create varied representations of the original text.

Following pre-processing, the data will be utilized to fine-tune the chosen model, a step that entails training the model on our specific dataset to tailor the pre-existing model weights for enhanced performance on our sentiment analysis task, focusing on detecting Plutchik's emotions. The fine-tuning phase will be carefully managed by selecting an appropriate learning rate and determining the optimal number of training epochs through initial experiments, aiming to achieve a balance between model accuracy and minimizing the risk of overfitting. We will also investigate the use of transfer learning techniques, such as domain adaptation, to leverage knowledge from related sentiment analysis tasks and improve the model's performance on our specific domain of olfactory experiences.

1.2.4 Model Evaluation. Evaluating the performance of our sentiment analysis model in detecting Plutchik's emotions presents a challenge due to the lack of a gold-standard dataset or ground truth labels for the Odeuropa dataset. To address this, we employ a combination of quantitative and qualitative evaluation methods, aiming to assess the model's effectiveness from multiple perspectives.

Firstly, we use cross-validation to assess the model's performance in detecting Plutchik's emotions across different subsets of the data. By splitting the dataset into multiple folds and iteratively training and testing the model on different combinations, we can obtain a more robust estimate of its generalization ability. This approach helps mitigate the potential bias introduced by evaluating the model on a single split of the data and provides a more comprehensive assessment of its performance.

Additionally, we conduct a manual analysis of a random sample of the model's predictions to gauge its effectiveness in capturing the specific emotions from Plutchik's wheel expressed in the olfactory descriptions. This qualitative evaluation involves domain experts examining the model's output and assessing the alignment between the predicted emotions and the underlying emotional tones conveyed in the text. By involving human judgment, we can gain valuable insights into the model's ability to capture the nuances of Plutchik's emotions and identify areas for further improvement.

To further validate our findings, we will explore the use of external datasets or resources that contain annotated examples of Plutchik's emotions in text. While these datasets may not be directly related to olfactory experiences, they can serve as a benchmark to assess our model's performance in detecting the eight primary emotions and their combinations. By comparing our model's predictions against these external references, we can gain additional confidence in its ability to accurately identify Plutchik's emotions.

Moreover, we will conduct error analysis to identify patterns in the model's misclassifications and gain insights into the challenges posed by the complexity of Plutchik's emotion model. This analysis will involve examining the specific instances where the model struggles to accurately predict the emotions and investigating the linguistic features or contextual factors that contribute to

these errors. By understanding the limitations of our model, we can develop targeted strategies for improvement, such as incorporating additional training data, refining the model architecture, or leveraging domain-specific knowledge to enhance its performance.

By combining these evaluation approaches, we aim to gain a comprehensive understanding of our BERT-based sentiment analysis model's performance and its suitability for unraveling the emotional dimensions based on Plutchik's wheel of emotions in the olfactory experiences captured in the Odeuropa dataset. Through rigorous quantitative and qualitative assessment, we strive to ensure the robustness and reliability of our findings, providing valuable insights into the complex emotional landscape associated with olfactory descriptions.

REFERENCES

[1] ROBERT PLUTCHIK. 1980. Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION. In *Theories of Emotion*, Robert Plutchik and Henry Kellerman (Eds.). Academic Press, 3–33. https://doi.org/10.1016/B978-0-12-558701-3.50007-7