Exploring the potential of automating the process of clustering smell stories

To Eun Kim, Sabina-Maria Mitroi, Rakshita Kumar, Neha Ranade, Karunya Selvarantnam, and John Xu
Department of Computer Science, University College London

***Abstract*—Experiences involving interaction with smell (olfactory experiences) have received less focus compared to auditory and visual experiences, even though they are more emotionally engaging and evoke memories that are autobiographically older. To delve deeper into this category of human experience, this paper explores how olfactory experiences captured by smell stories can be clustered using Natural Language Processing (NLP) and Machine Learning (ML). We take an iterative approach so that the NLP and ML techniques can be adjusted after obtaining the results from each iteration. The 3 key steps followed in every iteration are: Pre-Processing, Vectorising and Clustering. Different techniques are tried out for each of the 3 steps. To evaluate whether our clusters represent different kinds of smell experiences, we look for whether the clusters cover the 3 main components of smell stories - emotions, memories, and physical objects. We discover that sentiment-based vectors perform the best based on mathematical analysis whereas sentence vectors performs the best based on human analysis.**

1. Introduction

Smell is unique compared to other modalities (sight, sound, touch) because, memories evoked by smell give stronger feelings of going back in time, are more emotionally engaging, and are experienced more strongly [11], [39]. This ability of smell, to evoke a wide range of emotions and memories, makes understanding how smell affects human experiences an area of commercial value and research. For example, a better understanding of how smell influences mood can help interaction designers employ specific fragrances to improve user experiences.

One step towards understanding how smell affects experiences is to ask: *"What are the different ways in which smell shapes our experience?"*. Current research done to tackle this question involves analyzing hundreds of smell stories (short descriptions of experiences involving interactions with smell) to identify and categorize the different types of experiences stimulated by smell [26]. More specifically, due to the effect of smell on human experience, these smell stories often contain strong **emotions** and vivid accounts of **recalling memories associated with smells**. As expected, they also contain descriptions of **physical objects** such as *"bread, eggs, fuel"* that influence and often even initiate the experience. Further, analysis into these stories can enable us to understand the role smell plays in shaping human experience.

So far, techniques used to extract key categories of smell experiences from smell stories involve manual/physical categorization. For example, Obrist et al produced a paper [26] that manually groups smell experiences described in short smell stories into 10 broad categories. They found that these clusters provided an experienced-focused insight into smell stories that was then further used to link specific smell experiences to a potential applications in technology. For example, they identified that one group of smell stories were about how certain events in a person's life is associated with smell. They built further on these categories and conceptualised a remote sharing system where smells linked to a special event e.g. the smell of a Christmas dinner, could be sent to family and friends.

However, manually grouping the smell stories can be time consuming and can introduce biases [26]. Furthermore, Natural Language Processing (NLP) techniques have been developing rapidly making it possible to use Machine Learning (ML) algorithms to solve problems about textual data. Therefore, we aim to propose a method to automatically generate categories of smell experiences using NLP and ML techniques. Furthermore, since physical objects, emotions and memories (i.e. recalling memories), are 3 experience-related components frequently present in smell stories, we shall continuously reflect on these 3 elements and the role they play in defining and distinguishing clusters of smell stories.

The process of automating the categorisation process allows this paper to explore an alternate classification schema of smell stories that complements the work of Obrist et al. [26]. Beyond this, the clusters generated, aim to inspire further avenues of research exploring how smell experiences are expressed through language. Most importantly, the paper hopes to broaden our understanding of smell and human experience which can further the development of technologies where smell aids in improving user experience.

2. Related Work

This section explores the related work that has occurred in the areas of NLP, clustering and smell stories. Firstly, we look at why it is significant to cluster, identify and interpret smell stories based on experience. Secondly, we look into how these smell experiences are expressed in the English Language. And finally, we look at past work on clustering text, looking further into how text is represented through vectors.

2.1. Smell And Its Impact On Human Experience

The human experiences captured by smell stories often involve recalling the memories linked to a smell due to the strong link between memories and smell. de Bruijn, Maaike J. and Michael Bender [3] showed that with the aid of olfactory cues, participants were able to recall childhood memories, autobiographically older memories, much more vividly than with visual stimuli. Furthermore, Glachet, Ophélie, et al. [9] investigated the effect of odor exposure to retrieve memories in Alzheimer's patients, both recent and earlier memories. Alzheimer's disease patients recalled more memories after being exposed to smells than they did without exposure.

Smell also has a profound impact on emotions, which is another component that is frequently present in the experiences captured by smell stories. Helmefalk, Miralem, and Bertil Hultén [10] showed that smell can positively impact the emotions of potential consumers - increasing time and money spent. Olfactory cues, along with visual cues, impacted the behaviour and emotions of consumers. Villemure et al. [37] showed that focusing on odors reduced the unpleasantness of pain, and that odor altered moods and anxiety levels. Being exposed to negative odours induced negative moods such as increased anxiety level, while more pleasant odours resulted in an improved mood and a calming effect.

Physical objects are also an essential part of smell experiences. They can be the stimuli to many smell stories and the presence of these objects emphasise the smell detected. Verhulst et al. [36] investigate the presence of visual stimuli in Virtual Reality as a way of introducing smells. Participants reported that the presence of an image makes them think about smells, and can even elicit smells.

2.2. Smell And Its Representation In The English Language

Human experiences and changes to these experiences are communicated through language and the English language has evolved to describe our sensory experiences [40]. The language of smell has its own qualities.

Majid [19] explains that, unlike other senses such as tastes and textures, olfaction does not really have a dedicated lexical vocabulary. As a result, humans tend to describe words using source-based descriptors, such as *"smells like lavender"*. Xiao, Tait and Kang [41] explore how we perceive smells, finding that we describe smell through metaphors to express our emotions, feelings and evaluations of the environment. They found that humans mainly describe smells as these binary adjectives: fresh-stale, good-bad, happy-sad, harmonious-inharmonious, healthy-unhealthy, beautiful-ugly, smooth-rough, clean-dirty and safe-dangerous. They found that English speakers conveyed perceptual information such as what the smell was (identification), the source (location), their opinions (evaluation) when describing their smell experiences.

2.3. NLP Techniques: Clustering/Classification

Obrist et al [26] manually clustered smell stories into 10 categories, however, our paper aims to automate this process. Since, to the best of our knowledge, there is no attempt in the literature to automatically cluster smell experiences, we will summarize the most relevant studies that focused on (1) clustering/classifying stories by topic and (2) generating word-vectors to represent these stories.

Xu et al. [43] propose a DE-CNN convolutional neural network to be able to classify short texts, such as movie reviews, news and personality datasets. They focused on extracting the concepts of the text and then the context, and finally extracted the context-relevant-concepts which are then used to represent the text during the clustering of the texts. Abualigah et al. [2] proposed the idea that feature selection within text before clustering will result in increased performance of the clustering and reduces computation time. They used the particle swarm optimisation algorithm (computational method that optimizes a problem by iteratively trying to improve a candidate solution) to create features that are informative.

Lee and Jiang [15] proposed a fuzzy based method to classify texts. Since, they stated that one text can fall under multiple categories, they showed that relaxing the boundaries between the different categories, can allow more complex regions to be created, resulting in increased performance and speed. Selvi, S. Thamarai et al. [32] proposed a system where they use TF-IDF to represent how often a word appears in the document. They then found the cosine similarities between the given categories and the documents they were classifying. By filtering out the categories that have low cosine similarities, they passed the documents and their categories to supervised learning algorithms in order to be able to classify new unseen texts. Lin, Jiang and Lee [16] showed clustering is improved by computing the similarity between documents with respect to a feature, i.e. checking whether this document is composed of this feature, checking whether this feature belongs in both documents, only in one or in none. For cases where both are involved, the distance between the two becomes closer.

3. Research Aim

In this paper we aim to study how the olfactory experiences captured by smell stories can be automatically clustered. To achieve this:

1 - The paper endeavours to use clustering algorithms together with NLP to automate the process of clustering smell stories.

2 - The paper will simultaneously introduce different techniques (word vectors, extraction of nouns and adjectives from stories) and explore the extent to which the three elements of smell stories (physical objects, emotions, memories) have defined and distinguished the smell stories.

3 - The paper particularly focuses on exploring and identifying the vectorisation techniques that best represents the smell stories.

4. Research Design

4.1. Dataset description and collection

This study utilised 2 datasets: Base Dataset and Augmented Dataset.

4.1.1. Base Dataset. The data used in the preparation of this study was obtained from the paper by Obrist et al. [26], which, like this study, aimed to generate clusters of smell stories. The paper collected smell stories, the title for these smell stories and other information such as when the story took place, how positive or negative the experience was. from 439 participants. This study only utilises the textual part of the dataset consisting of the 439 smell stories and the corresponding story titles.

4.1.2. Augmented dataset. Most of the state-of-the-art NLP models must be trained on a large corpus. Therefore, it seems insufficient to train NLP models from scratch with the dataset of 439 stories.

Out of many text augmentation methodologies, we chose three different methods (Data Noising, Lexical Substitution and Back-translation) and chained those together into a pipeline. These 3 key stages have been described below.

Data Noising: For the first step, random noise was injected into our text data either by insertion, swapping or deletion. For insertion, the algorithm randomly chose a non-stop word from the Wordnet Thesaurus [22] and this was inserted into the text at a random position. For swapping, the algorithm chose any two words from the text and swapped their positions. Lastly, for deletion, the algorithm deleted one random word from the text [38].

Lexical Substitution: The data resulted from the Data Noising stage is used as input for lexical substitution. This research used thesaurus and word embedding-based substitution [38] to transform our text by replacing some of the words with its synonym.

Back Translation: Back translation refers to a procedure of translating a text from language *"A"* to another language *"B"* and then translating back into language *"A"* to get transformed text which is syntactically different but semantically identical [42]. To get the most promising result from the English text corpus, Urdu, Vietnamese and Tamil were chosen as they are linguistically far from English.
In the end we have two datasets:

1) **Base dataset:** containing the original 439 stories which are finally clustered by the clustering algorithms.

2) **Augmented dataset:** containing 1756 stories used for the purpose of training the NLP models.

4.2. Methodology

In this paper we aim to use NLP and ML techniques to study how the experiences captured by smell stories can be clustered. The process of transforming the stories into their respective clusters involves 3 key steps: Pre-Processing, Vectorisation and Clustering.

An **iterative approach** (summarised in Figure 1) was adopted to enable an exploration of a wide range of technologies and techniques across these three steps. This section provides an overview of the technologies and techniques covered by each of the three key steps.

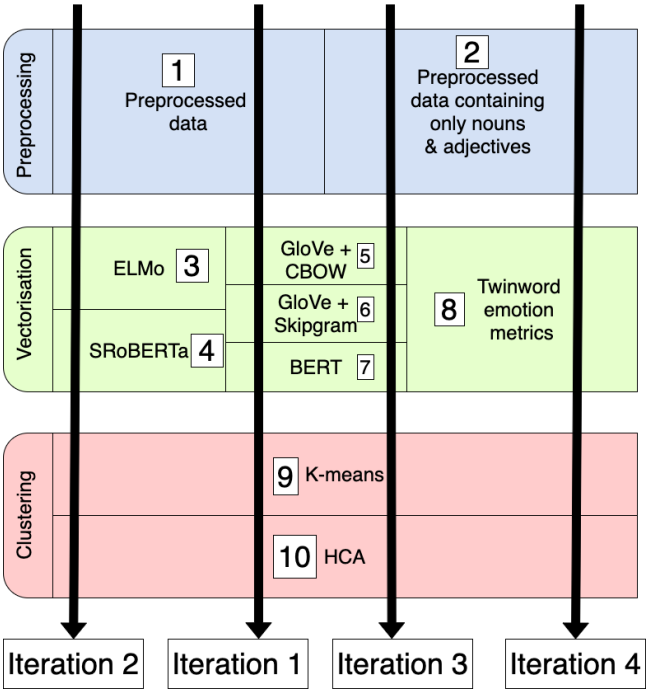

Figure 1: *Summary of Iterations.* Follow the arrows to understand what technologies and techniques were used across 3 steps for each of the 4 iterations.

4.2.1. Pre-processing Smell Stories. It eliminates noise from the text and extracts key information that improves the clustering quality [33]. First, we appended each smell story with the user-inputted titles to enrich the existing data. Then, punctuation, accents and stopwords were removed from the stories since they add noise. Each story was tokenized and the text lowercased. This Pre-Processing pipeline is depicted by *"1"* in Figure 1. Later iterations (iteration 3 and 4) further adapted the Pre-Processing pipeline to remove all words apart from nouns and adjectives which generated a second dataset (as shown by *"2"* in Figure 1).

4.2.2. Vectorising Smell Stories. Clustering algorithms work with numerical data and cannot understand textual data the way humans do. Therefore, the main motive of this stage is to generate vectors that effectively capture and distinguish the experiences described in smell stories for clustering. This section looks at the key ways we vectorise smell stories.

Word embeddings/vectors incorporate context to generate vectors for each word in the text. These vectors can be combined to form vectors for the smell stories. The word embedding models utilised were:

1) **Word2Vec & GloVe**: Word2Vec consists of a family of word embedding techniques that use neural networks to generate word-vectors [1]. The two key Word2Vec algorithms are CBOW and skip-gram. GloVe is an *"unsupervised learning algorithm for obtaining the vector representation of words"* [27] that provides a set of pre-trained vectors. To generate word-vectors that better capture the language domain of smell stories, the skip-gram and CBOW models finetuned the pre-trained GloVe vectors on the augmented dataset. (shown by *"5"* and *"6"* in Figure 1).
2) **BERT**: BERT is a *"neural network-based technique"* that reached state-of-the-art results in a number of tasks using transformer architecture [7], [24]. Our paper trains BERT on the augmented dataset of smell stories to generate word-vectors (shown by *"7"* in Figure 1).

Once the word-vectors had been created via the above three techniques, they were combined by averaging each word vector in the story (Iteration 1, Figure 1) or by averaging only the vectors for the nouns and adjectives in the text (Iteration 3, Figure 1).

Sentence embeddings/vectors incorporate the entire sentence when generating the vectors which enables the nuances in the context of the words to be considered [29]. For example, here the meaning of bucket is different in these two sentences : *"This is my bucket list"*, *"The bucket is full"*.

The sentence embedding models this study uses are: Sentence RoBERTa (SRoBERTa) [29] and ELMo [28]. RoBERTa is an improved version of BERT (trained for more hours on a larger corpus). [18]. ELMo, a character-based double-layered bidirectional LSTM model, is able to generate multiple word-vectors in accordance with its context. Lower layers tend to capture the syntax while higher layers capture semantics. [28]. SRoBERTa (pre-trained for Semantic Textual Similarity (STS) and then finetuned on the augmented data) and ELMo were used to generate a set of sentence vectors for the smell stories (*"3"* and *"4"* in Figure 1) .

Apart from word and sentence embeddings, this study explores **Sentiment-Based Vectors**. Due to the strong link between smells and emotions - sentiment-based vectors emerged as a vectorisation technique.

Plutchik's Wheel of Emotions states there are 8 basic emotions [28]. To generate the vectors based on these emotions, the Twinworld API was used (*"8"* in Figure 1), which captures 6 emotions (*'joy'*, *'sadness'*, *'surprise'*, *'fear'*, *'anger'*, *'disgust'*). This was chosen over alternative libraries and APIs that captured a fewer number of emotions.

4.2.3. Clustering the smell stories. Once the text is converted to vectors, clustering algorithms can be applied to group the stories into clusters. This section covers the clustering algorithms used by this study.

K-means algorithm is an iterative algorithm that partitions the dataset into K pre-defined distinct subgroups where each data point belongs to one group [14]. This is one of the clustering techniques used by our paper (*"9"* in Figure 1).

Hierarchical clustering is a general family of clustering algorithms that builds nested clusters by merging or splitting them successively [12]. This is also one of the clustering techniques used by our paper (*"10"* in Figure 1).

4.3. Metrics

The metrics used to analyse the quality of the clusters generated can be divided into 2 parts.

4.3.1. Mathematical Metrics. In this project, we did not possess pre-defined cluster labels for the stories so, 3 different internal clustering validation metrics were chosen: Silhouette Scores (S score), Calinski-Harabasz Index (CH Index), and Davies-Bouldin Index (DB Index).

As the goal of clustering is to make objects within the same cluster similar and objects in different clusters distinct, internal validation measures are often based on the following two criteria [35] [45]: (1) **Compactness**: It measures how closely related the vectors in a cluster are. (2) **Separation**: It measures how distinct or well-separated a cluster is from other clusters.

Ideally, the more compact and distinct the cluster is, the higher the Silhouette coefficient and Calinski-Harabas index scores are and lower the Davies-Bouldin index is. We use all three metrics to evaluate compactness and separation for better accuracy and robustness. The details of these three mathematical metrics can be found in Appendix section.

4.3.2. Human-based Analysis. There were 3 main components of our human-based analysis:

1) **What does each cluster represent:** We assigned category labels to the clusters by looking at the top 5-10 most frequently occurring words in the word-cloud. This allowed us to evaluate whether the clusters produced, separated out the smell stories based on a certain commonality. To correctly assign these labels, we generated word clouds for each cluster. The generated labels were then cross-checked by looking into the actual stories within each cluster. A word cloud and its corresponding label identified after human analysis is shown in Figure 2 below.
2) **Distinctiveness of each cluster:** We looked for overlaps and the distinctiveness in the category labels identified in the previous point. This allowed us to verify our results from the mathematical metrics.
3) **Identifying the 3 components of smell stories in the cluster labels:** We analysed the labels provided by each cluster and tried to understand the role played by the 3 components of smell stories (emotions, physical attributes and memories) in defining and distinguishing clusters.

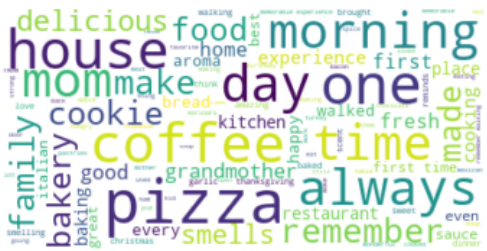

Figure 2: *ELMo (Hierarcjical Custering).* Cluster 3 - Home-food related experiences

5. Analysis of Results

Our results for each iteration can be found in Figure 3. Through the 4 iterations, we attempted to improve the quality of clustering by adjusting the techniques used,

5.1. Iteration 1: Word vectorisation

In the first iteration, we used word-vectorisation techniques which included Skip-gram, CBOW (both of which were pre-trained by using vectors generated by GloVe) and BERT [vectorisation stage]. Alongside this, the pre-processed dataset was used [preprocessing stage] and K-means and Hierarchical Clustering were used for clustering [clustering stage].

5.1.1. Mathematical Analysis. From the mathematical metrics in Figure 3, we can see that CBOW outperforms Skip-gram and BERT, suggesting that clusters generated after using CBOW are more distinct and compact. A possible reason for this higher performance is rooted in the fact that CBOW provides better representation for frequent words [23]. We are dealing with smell stories and this is a very specific type of text we are looking at. This means there is a high frequency of certain words particularly *"remember"*, *"smell"*, and important smell descriptors e.g. *"rotten"*. Given these patterns, CBOW provides a better representation of important words which can contribute to a better representation for the overall story.

Furthermore, all metrics for the CBOW vectorisation technique indicate that K-means performed better than hierarchical clustering. This perhaps shows that the ideal clusters of the vectors resemble a hyper-spherical shape - a characteristic which allows K-means to excel. [13].

5.1.2. Human Analysis. After human analysis was carried out, we labelled the clusters and generated the following categories for each method:

- *Skipgram* : [Cluster 0] Negative emotions [Cluster 1] Food experiences [Cluster 2] Recalling smell linked memories [Cluster 3] Pleasant atmosphere
- *CBOW*: [Cluster 0] Positive emotions [Cluster 1] Pleasant experiences [Cluster 2] Food experiences [Cluster 3] Undefined
- *BERT*: [Cluster 0] Homely/Comforting Memories [Cluster 1] Undefined [Cluster 2] Undefined

Firstly, we will explore the distinctiveness of clusters. As expected based on the results of the mathematical metrics, BERT clusters are not distinct as we cannot observe a theme in the word-clouds for cluster 1 to 2. However, unlike the results from the mathematical metrics, we can observe that clusters generated using Skip-Gram as the vectorisation technique are more distinct than those generated by CBOW. Unfortunately, the discrepancy in results might be because of the limitations of our human evaluation techniques (explored in Limitations section).

Secondly, we observe that several of the labels have a strong link to one or more of three experiential components of smell stories (physical objects, emotions, memories). For example CBOW cluster 0 links to positive emotions (emotions), Skip-Gram cluster 2 links to Recalling smell linked memories (memories), Skip-Gram cluster 1 links to food experiences (physical objects).

5.1.3. Evaluation. Overall, our best results are obtained by CBOW and Skip-Gram where the former performs better in numerical metrics and latter performs better in human-based metrics. Despite this the mathematical metrics for both are close to 0 indicating that the clusters are not compact or distinct. The reason for this could be because we averaged the word-vectors to generate a representation of each story. Although the technique of averaging the word-vectors to generate sentence vectors is known to be computationally faster than other techniques, it is also known to worsen the quality of vectors, thereby affecting the clusters [44]. In hopes of further improving on this shortcoming, we explored methods in which we can directly generate vectors for the text without needing to combine vectors via averaging.

5.2. Iteration 2: Sentence vectorisation

Improving upon the limitations of techniques used in Iteration 1, this section covers the results from the sentence vectorisation techniques: ELMo and SRoBERTa [vectorisation stage]. Like Iteration 1, the pre-processed dataset was used [preprocessing stage] and K-means and Hierarchical Clustering were used for clustering [clustering stage].

5.2.1. Mathematical analysis. From the mathematical metrics in Figure 3, it can be seen that there is a slight increase in score when SRoBERTa is compared to BERT in the Silhouette score of Hierarchical Clustering and the Calinski-Harabasz Index in K-means. However, apart from this, both SRoBERTa and ELMo performed poorly in all the other mathematical metrics compared to Iteration 1 since the Silhouette scores and Calinski-Harabasz indices decreased, while Davies-Bouldin indices increased. However, this can be explained by the nature of smell stories. It is natural that one smell story can be assigned to multiple clusters since they all have a commonality: they all involve descriptions where smell is at the centre. Therefore, if sentence vectors represent and capture the meaning of the stories well

	Word Vectors Iteration 1			Sentence Vectors Iteration 2		Enhanced Pre-Processing Iteration 3		Sentiment Vectors Iteration 4
	Pret-Trained Skip-Gram	Pre-Trained CBOW	BERT	ELMo	SRoBERTa	Pre-Trained Skip-Gram	Pre-Trained CBOW	Sentiment Vectors
S Score (HCA)	0.054	0.209	0.033	0.028	0.046	0.088	0.048	**0.526**
S Score (K-Means)	0.074	0.176	0.071	0.043	0.039	0.101	0.075	**0.565**
CH Index (HCA)	31.519	93.745	23.122	13.232	18.052	46.426	25.6	**446.732**
CH Index (K-Means)	31.757	104.873	31.566	17.225	42.716	53.258	29.926	**542.577**
DB Index (HCA)	3.261	2.07	3.234	4.071	3.977	2.844	2.954	**0.742**
DB Index (K-Means)	3.141	2.053	2.732	3.44	3.82	2.859	2.778	**0.729**

Figure 3: *Overview of the S Score (Silhouette Score), CH Index (Calinski-Harabas Index) and (Davies-Bouldin Index) DB Index across the 4 iterations.* For a more detailed version, check here: 20

enough, they can, ironically, perform poorly in K-means and Hierarchical Clustering evaluation metrics since all the stories are similar. Since the mathematical metrics measure the distinctiveness and compactness of clusters, this can produce low scores.

5.2.2. Human analysis. From the generated word clouds, we manually labelled each cluster as below:

- *ELMo:* [Cluster 0] Indoor-related memories [Cluster 1] Nature Memories [Cluster 2] Pleasant experiences [Cluster 3] Food experiences [Cluster 4] Unpleasant experiences
- *SRoBERTa:* [Cluster 0] Recalling smell linked memories [Cluster 1] Food experiences [Cluster 2] Pleasant experiences [Cluster 3] Strong Unpleasant experiences

Let us first analyse the distinctiveness of the clusters. Although the mathematical scores didn't seem promising, the clustering results for both ELMo and SRoBERTa visualised by word clouds are distinct. For vectorisation with SRoBERTa we have four different clusters with little overlap between them. For example cluster 0 is memory based, cluster 2 is about pleasant experiences while cluster 3 is about unpleasant experiences, and cluster 1 is food related. Similarly the clusters made by SRoBERTa are also very distinctive and could be easily labelled. For example, in cluster 1, indoor-related smell experiences are captured while in cluster 0, nature and outdoor memories are captured. These results can be contrasted with Iteration 1 in which several clusters are "undefined" and difficult to label due to the mix of different types of stories.

All three components of smell stories (emotions, physical objects, and memories) are also present in the cluster labels generated. Cluster number 0 and 1 from ELMo is majorly linked to memories, cluster 3 is linked to the physical objects, and cluster 4 focuses on negative emotions. Similar observations can be seen by SRoBERTa.

5.2.3. Evaluation. The overall results demonstrate that sentence vectors generated from bidirectional language models can result in poor mathematical scores when it comes to clustering. However, as can be seen from the human analysis, the clusters generated are easily distinguishable and effectively cover the 3 components of smell stories.

Overall, a limitation of iteration 2 is that the clusters are not distinct or compact (indicated by the low maths scores). To fix this issue, the distinctiveness could be improved between the stories by passing only nouns and adjectives to the vectorisation stage. This could highlight some important elements in the stories that can better define and distinguish the texts. We have explored this modification in the next iteration.

5.3. Iteration 3: Enhanced pre-processing

Unlike the previous iterations, in this section we analyse the results obtained by utilising the dataset containing only nouns and adjectives [preprocessing stage]. The vectorisation method utilised were the Word2Vec models from iteration 1 (Sentence vectorisation could not be included because they require full sentences) [vectorisation stage]. Again, the produced embeddings were clustered using K-means and Hierarchical Clustering [clustering stage].

5.3.1. Mathematical analysis. Compared to Iteration 1, for CBOW, the Davies-Bouldin score increases and the Silhouette and Calinski-Harabasz score goes down, indicating a worser level of clustering. At the same time, when compared to Iteration 1, for Skip-Gram, the Davies-Bouldin score reduces and the Silhouette and Calinski-Harabasz score increases, indicating an improved level of clustering. These results indicate that the use of the nouns and adjectives

dataset lowered the performance with CBOW and improved the performance with Skip-gram - although only marginally. A possible explanation for this reduction in performance is rooted in the key differences between Skip-Gram and CBOW. While Skip-Gram works well with small amounts of data, CBOW has better performance for larger amounts of data [21]. Overall the number of words reduced by 62.5% after all words apart from nouns and adjectives were removed. This reduction in data may have led to a worser performance for CBOW. Like the previous iterations, the clustering performance is higher with k-means. This perhaps indicates that even after the pre-processing pipeline was modified, the ideal cluster shapes for the data is hyper-spherical [13].

5.3.2. Human analysis. After looking at the word clouds, we were able to categorize the clusters with the following labels:

- *Skip-gram:* [Cluster 0] Strong emotional memories, [Cluster 1] Food experiences, [Cluster 2] Positive emotional memories
- *CBOW:* [Cluster 0] Mixed first emotions, [Cluster 1] First time with food experiences, [Cluster 2] Food experiences, [Cluster 3] Positive first-time emotions

Firstly, looking at cluster labels generated through human-analysis for both Skip-gram and CBOW, we can argue that CBOW labels are less distinct and more overlapping. This is supported by the fact that there are two overall categories (emotions and food) that contain overlapping subcategories for CBOW since cluster 1 and 2 are food related while cluster 0 and 3 are emotion linked. However in Skip-gram only cluster 0 and 2 are linked by the fact that they both describe memories. This presence of overlaps potentially explains why CBOW performed poorly on numerical metrics compared to Skip-gram.

Secondly, the cluster labels cover the 3 different components of the smell stories (emotions, memories and physical entities). For example, CBOW cluster 3 and Skip-gram cluster 0 links to positive feelings (emotion), CBOW cluster 2 and Skip-gram cluster 1 is related to food items (physical entities), Skip-gram cluster 2 is related to recalling memories (memories). Therefore, like Iteration 1, the clusters capture all the 3 components of smell stories.

5.3.3. Evaluation. Overall, comparing the results of this iteration to the results of the first iteration, we can see that the current iteration gives poorer separation between clusters according to the mathematical metrics. This might be because by picking only nouns and adjectives, we might have made the dataset narrowly focused without capturing any context. This context is important for Word2Vec techniques to effectively capture the text [21].

Finally, the lower result we experienced indicated another clear direction for our investigation. Since sentence and word vectors did not improve the mathematical scores, an alternate solution was to explore a new vectorisation technique. The strong link between smell and emotions encouraged the exploration of a vectorisation technique focused on emotions. This technique is explored in the next iteration.

5.4. Iteration 4: Sentiment-based Vectors

In this section we analyse the results obtained by utilising the dataset containing only nouns and adjectives [pre-processing stage] to generate the sentiment-based vectors via the TwinWorld API [vectorisation stage], which was finally clustered using K-means and Hierarchical Clustering [clustering stage].

5.4.1. Mathematical analysis. Comparing the sentiment-based vectors with other results, it can clearly be observed that there is a stark improvement in the performance. This time there is an average 183%, 398%, -64% change in Silhouette, Calinski-Harabasz and Davies-Bouldin scores respectively compared to the best mathematical results we have received so far (Iteration 1 - CBOW). This observation further supports the strong connection between smell and emotions as described in published literature [10], [37]. Interestingly the nouns and adjective dataset improved the results for sentiment-based vectors and worsened the results for word-vectors (as shown by Iteration 3 in Figure 3). This observation can be explained by the differences in word and sentiment-based vectorisation: isolating the adjectives and nouns can remove the context needed to generate word vectors [21], however for sentiment-based vectorisation it removes noise since adjectives and nouns are among the important parts-of-speech in determining sentiment [25].

Like the previous iterations, the clustering performance is higher with K-means. This perhaps indicates that even after sentiment analysis is used, the ideal cluster shape for the data is hyper-spherical [13].

5.4.2. Human analysis. We have these labels for the clusters generated for this iteration:[Cluster 0] Positive emotions, [Cluster 1] First time emotions, [Cluster 2] Negative emotions.

Looking at the distinctiveness between the clusters, the distinguishing factors between cluster 0 and 2 is clear. Cluster 0 is associated with positivity while cluster 2 is associated with negativity. Further analysis of the emotions within each cluster can support this distinction between cluster 0 and 2. The 6 emotion metrics across the stories in each cluster were averaged producing Figure 4 below.

The graph clearly shows the cluster 0 is linked to positive emotions (i.e. Happy) while cluster 2 is linked to negative emotions (i.e. Fear, Angry Disgust, Sad)

However for cluster 1 the table in Figure 4 shows the distinction between the different emotions is less clear. Nevertheless, there still appears to be a distinction since the top words for this cluster (*"one, fresh, scent, perfume"*) do not seem to overlap with the top words of the other clusters. Overall, the higher level of distinctiveness across all cluster supports the improved scores for the mathematical metrics compared to previous iterations.

	SAD	HAPPY	SURPRISE	FEAR	ANGRY	DISGUST
Cluster 0	0.001	**0.419**	0.061	0.002	0	0
Cluster 1	0.016	0.054	0.031	0.013	0.018	0.013
Cluster 2	**0.124**	0.003	0.045	**0.159**	**0.135**	**0.149**

Figure 4: *Iteration 4: Average Emotions for Each Category Generated.* The bold values highlight the emotions that dominate for each cluster. For a more detailed version, check here: 21

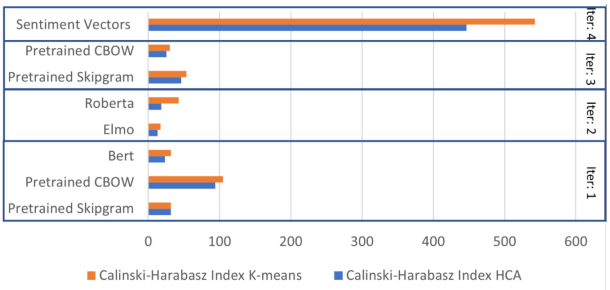

Figure 5: *Calinski-Harabasz Index Over 4 Iterations*

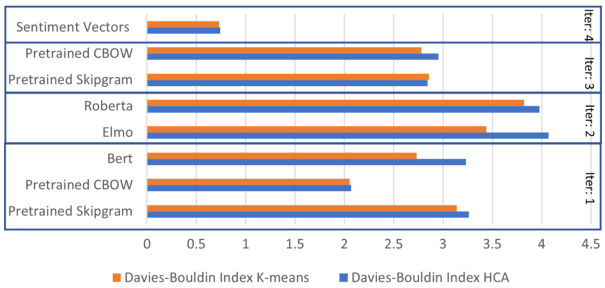

Figure 6: *Davies-Bouldin Index Over 4 Iterations*

Although the clusters are distinct, this iteration uses sentiment-based vectors which focuses more on the emotion behind the smell stories instead of focusing on all three components of smell stories. This may lead to certain components of the smell stories to be overlooked. This can clearly be seen by looking at the cluster labels. The clusters generated by sentiment-based vectors are focused more on emotions: *"positive emotions"* (cluster 0) and *"negative emotions"* (cluster 2). However in previous iterations the clusters generated were linked to all three components of smell stories: *"Food experiences"* (physical memories), *"Recalling smell linked memories"* (memories), *"Positive emotions"* (emotions).

5.4.3. Evaluation. The overall results show that sentiment-based vectorisation appears to improve the clustering according to mathematical metrics.

Although a similar improvement appears to exist according to human analysis, the clustering focuses mainly on emotions. Perhaps generating a system to combine sentiment-based vectors with word/sentence vectors could improve the cluster generate and ensure it includes all the key components of a smell story.

6. Discussion & Limitations

In this paper we aimed to use NLP and ML techniques to study how the experiences represented in smell stories can be clustered. Below, we discuss how well our results have fulfilled the aim by comparing the results from the clusters that performed best in terms of the mathematical and human analysis that was carried out.

A summary of the key trends observed in mathematical metrics over the 4 iterations can be visualised in Figure 5, 6 and 7.

According to figure 5, 6 and 7 , the best mathematical results arise from utilising sentiment-based vectors to represent the smell stories. An important question to ask here is *"Why do sentiment-based vectors outperform other methods in terms of the mathematical clustering?"* The sentiment-vectors do not try and represent everything about the stories, but only the emotions present in the vector (noise-free). Therefore, they have lesser noise and are of fewer dimensions (6 vs 100) [6]. This narrow focus of sentiment-based vectors makes them easier to cluster, hence the high mathematical performance.

Returning to the aim of the paper, it is worth mentioning that emotions are an important component of most smell experiences, as mentioned in the background research [3] [9]. Therefore, high performance of sentiment-based vectors means that we are able to successfully cluster smell stories based on an important component of experience, emotions.

At the same time, sentiment-based vectors do not perform the best according to human-analysis. Although sentiment-based vectors formed 3 well-defined and distinct categories, they are limited in their ability to represent all components that make up a smell experience captured by a smell story (besides emotions) [31].

Therefore although the sentiment-based vectorisation produces the best clusters, it does not entirely fulfill the aim - To cluster smell stories according to all components that make smell experience. Smell stories are defined to be much broader than just emotions.

Sentence vectors on the other hand, entirely fulfills our aim and shows the best performance according to the human-analysis results shown by figures 8-11 (Appendix Section). An important question to ask here is *"Why do sentence vectors perform well in human analysis?"* Sentence vectors that we have used, go to great lengths to capture the entire context of the stories. The polymorphic nature of word sentence vectors, character-level network in ELMo and Semantic Textual Similarity optimization in SRoBERTa together contribute to better capturing the essence of smell experiences, compared to any other technique [28].

However, they perform quite poorly in terms of the mathematical metrics. This is expected: experiences are

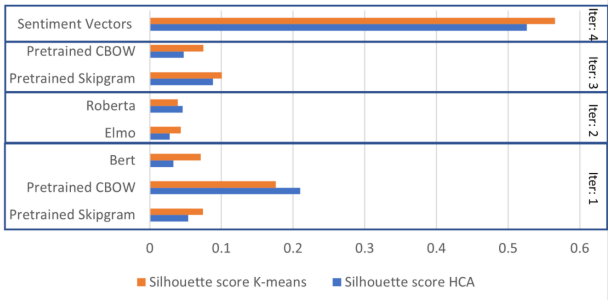

Figure 7: *Silhouette Scores Over 4 Iterations*

not clearly distinguishable into separate areas, rather, it is made up of overlaps between emotions, memories and physical objects [8]. The lower performance in clustering (the mathematical metrics) tells us that clustering is taking place, however, there exists overlaps and the clusters are not distinct.

Overall, in terms of best vectorisation technique, we found that sentence vectors present the greatest potential in fulfilling the aims of this research paper, capturing a wider category of experiences from smell stories. At the same time, we found that sentiment-based vectors give us the most compact and separable clusters. Lastly, in terms of best clustering technique, throughout the iterations, K-means constantly performed better than Hierarchical Clustering, albeit marginally.

6.1. Limitations

Limitations of our methodology: In this paper, we focused on different vectorisation methods that would capture the essence of the smell stories, from word-vectors to sentence vectors and sentiment-based vectors. However, this is just one part of the process. Investigating other clustering algorithms could also reveal better automated processes to cluster smell stories.

During our human analysis we looked at the word clouds to decide what experiences they fall under. Although we get a good idea of the most common words that are used within the cluster - we do not get a good representation of the context with which the words themselves are used. The context could give us more information about the emotional context and even the time context.

Additionally, we used pre-trained language models to account for our lack of data. Vectors from pre-trained models were trained on a larger domain of the English Language, which incompletely represents the frequency and context in which smell language is used. This, then results in the vectors representing the stories to be near each other in the vector space, as semantically similar sentences would have vectors that are more similar (i.e. higher semantic relatedness) [34]. We are not able to completely distinguish these smell stories, resulting in overlaps, in other words, reduced cluster quality.

Lastly, since this is a novel area, and this dataset is unique, and previous work has not looked into the classification of smell based on experiences, we do not have a benchmark to check the cluster results against. Obrist et al. [26] manually clusters smell stories. However, it is not possible for us to use these as comparison. This is because we try to generate clusters that are fewer and try to capture the elements of smell experiences whereas, Obrist et al [26] divides experiences into more abstract categories, such as *"Social interaction is affected by the smell"*.

Limitations of our dataset: There are also specific limitations that originated from the dataset utilised in this study. The participants used to make the dataset were mainly from the American population [26]. Given the importance of culture in affecting our past experience and the way we experience the future, these limitations may have introduced a cultural bias within our results [20].

7. Conclusion and Future Work

Exploring the link between smell and human experiences is an important yet novel area of research. This paper provides an insight into this relationship by utilizing the NLP and machine learning algorithms to cluster smell stories. We take an iterative approach to conduct the experiment. After each iteration, we analyze the quality of the outcome through both mathematical and human-experience-based metrics. We then improve our machine learning and NLP models according to the analysis results before we start the next iteration of the experiment. After all iterations, we found that sentiment-based vectors give the most well-separated categories whereas sentence vectors give a set of clusters that better covers all 3 components of smell stories.

There are several promising directions for further research, arising from this literature.

The first area of future work arises from the vectorisation techniques that we have utilised. Although sentiment-based vectors performed well, they only included the emotions of the smell story and focused less on the other 2 components of smell stories e.g. physical objects, memories in the stories. Therefore, refining sentence vectors to include sentiment/emotion analysis could be a future area of work. Aside from this we can incorporate existing techniques for vectorisation (e.g. Doc2Vec and Top2Vec) and explore their effectiveness for representing smell stories.

Another area of future work arises from the preprocessing techniques we have utilised. In the future we could apply supervised machine learning to develop a model that helps extract smell words. This can allow us to cluster based on the sections of the text relevant to the user's smell experience perhaps leading to better results. Furthermore, this would provide a useful dataset that can firstly benefit this research and can also be used to expand this research area in several directions.

Additionally in the future, this research could be used to facilitate the development of smell technologies e.g. smell-based text-tagging, smell sharing on social media, smell based notification system. For example the clusters containing pleasant stories could be used by UX designers to identify how user experience can be improved.

8. References

[1] Word2vec nbsp;: nbsp; tensorflow core.

[2] Laith Mohammad Abualigah, Ahamad Tajudin Khader, and Essam Said Hanandeh. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science*, 25:456–466, 2018.

[3] Maaike Bruijn and Michael Bender. Olfactory cues are more effective than visual cues in experimentally triggering autobiographical memories. *Memory*, 26, 10 2017.

[4] Tadeusz Caliński and Harabasz JA. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 01 1974.

[5] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[6] DeepAI. Curse of dimensionality, May 2019.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] The Interaction Design Foundation. The classic types of experience, March 2016.

[9] Ophélie Glachet, Ahmed Moustafa, Karim Gallouj, and Mohamad El Haj. Smell your memories: Positive effect of odor exposure on recent and remote autobiographical memories in alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, 41:1–10, 03 2019.

[10] Miralem Helmefalk and Bertil Hultén. Multi-sensory congruent cues in designing retail store atmosphere: Effects on shoppers' emotions and purchase behavior. *Journal of Retailing and Consumer Services*, 38:1–11, 2017.

[11] Rachel S Herz and Trygg Engen. Odor memory: Review and analysis. *Psychonomic Bulletin & Review*, 3(3):300–313, 1996.

[12] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

[13] Manju Kaushik and Bhawana Mathur. Comparative study of k-means and hierarchical clustering techniques. *International journal of software and hardware research in engineering*, 2(6):93–98, 2014.

[14] K Krishna and M Narasimha Murty. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, 1999.

[15] S. Lee and Jung-Yi Jiang. Multilabel text categorization based on fuzzy relevance clustering. *IEEE Transactions on Fuzzy Systems*, 22:1457–1471, 2014.

[16] Yung-Shen Lin, Yi Jiang, and Shie-Jue Lee. A similarity measure for text classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 26:1575–1590, 07 2014.

[17] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. pages 911–916, 12 2010.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[19] Asifa Majid. Cultural factors shape olfactory language. *Trends in Cognitive Sciences*, 19, 10 2015.

[20] Asifa Majid and Stephen C Levinson. The senses in language and culture. *The Senses and Society*, 6(1):5–18, 2011.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[22] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[23] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112:340–349, 2017.

[24] Pandu Nayak. Understanding searches better than ever before, Oct 2019.

[25] Chris Nicholls and Fei Song. Improving sentiment analysis with part-of-speech weighting. In *2009 International Conference on Machine Learning and Cybernetics*, volume 3, pages 1592–1597. IEEE, 2009.

[26] Marianna Obrist, Alexandre Tuch, and Kasper Hornbæk. Opportunities for odor: Experiences with smell and implications for technology. *Conference on Human Factors in Computing Systems - Proceedings*, 04 2014.

[27] Jeffrey Pennington.

[28] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[29] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019.

[30] Peter Rousseeuw. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987.

[31] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[32] Thamarai Selvi Somasundaram, P. Karthikeyan, A. Vincent, V. Abinaya, G. Neeraja, and R. Deepika. Text categorization using rocchio algorithm and random forest algorithm. pages 7–12, 01 2017.

[33] V Srividhya and R Anitha. Evaluating preprocessing techniques in text categorization. *International journal of computer science and application*, 47(11):49–51, 2010.

[34] Mohamed Ali Hadj Taieb, Torsten Zesch, and Mohamed Ben Aouicha. A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelli-*

gence Review, 53(6):4407–4448, 2020.

[35] Tan, Pang-Ning, Michael Steinbach, Michael Adeyeye Oshin, Vipin Kumar, and Vipin. *Introduction to Data Mining*. 05 2005.

[36] Adrien Verhulst, Eulalie Verhulst, Minori Manabe, Hiroto Saito, Shunichi Kasahara, and Masahiko Inami. Investigating the influence of odors visuals representations on the sense of smell, a pilot study. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 727–728. IEEE, 2020.

[37] Chantal Villemure, Burton Slotnick, and Mary Bushnell. Effects of odors on pain perception: Deciphering the roles of emotion and attention. *Pain*, 106:101–8, 12 2003.

[38] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[39] Johan Willander and Maria Larsson. Smell your way back to childhood: Autobiographical odor memory. *Psychonomic bulletin & review*, 13(2):240–244, 2006.

[40] Bodo Winter, Marcus Perlman, and Asifa Majid. Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, 179:213–220, 2018.

[41] Jieling Xiao, Malcolm Tait, and Jian Kang. Understanding smellscapes: Sense-making of smell-triggered emotions in place. *Emotion, Space and Society*, 37:100710, 2020.

[42] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training, 2020.

[43] Jingyun Xu, Yi Cai, Xin Wu, Xue Lei, Qingbao Huang, Ho fung Leung, and Qing Li. Incorporating context-relevant concepts into convolutional neural networks for short text classification. *Neurocomputing*, 386:42–53, 2020.

[44] Xiaoya Yin, Wu Zhang, Wenhao Zhu, Shuang Liu, and Tengjun Yao. Improving sentence representations via component focusing. *Applied Sciences*, 10(3):958, 2020.

[45] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. *International Conference on Information and Knowledge Management, Proceedings*, 08 2002.

9. Appendix

The link to our code and files can be found here for a deeper understanding about our work: Click here.

Silhouette Coefficient: It is also known as Silhouette score(S) of a single point [17], is measured as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

In the given formula, $a(i)$ is the average intra-cluster distance of i^{th} object to all the other ones in the same cluster and $b(i)$ is the average inter-cluster distance of i^{th} in the closest cluster. The number $s(i)$ is obtained by combining $a(i)$ and $b(i)$ as follows:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i), \\ 0 & \text{if } a(i) = b(i), \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i). \end{cases}$$

This has been introduced by Rousseeuw et al. [30] in 1986. The Silhouette Coefficient takes into account both validation criteria and tells us how well-assigned each individual point is. This index falls into the range of $[-1, 1]$. If S is close to 0, it is right at the inflection point between two clusters. A value closer to -1 indicates that clustering configuration may have too many or too few clusters. A value closer to 1 indicates that points are well-assigned to its own cluster.

Calinski-Harabasz Index: It is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters [17]. Higher value of Calinski-Harabasz index indicates that the clusters are denser and well-separated. This has been introduced by Calinski et al. [4] in 1974. This can be used to assess the model when round truth labes are not known.

The Calinski-Harabasz for K numbers of clusters on a data set D which is $[d_1, d_2, ..., d_n]$ is presented as:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k \left\| c_k - c \right\|^2}{K - 1} \right] / \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| d_i - c_k \right\|^2}{N - K}$$

By explaining it in simple terms, n_k and c_k represent the number of point and the centroid of the k^{th} cluster. The global centroid is c and N is considered the total number of data points.

Davies-Bouldin Index: It takes only one validation criterion into account - separation [17]. This has been introduced by Davies et al. [5] in 1979.

The Davies-Bouldin index is calculated as follows:

1) For each cluster C, the similarities between C and all other clusters are computed, and the highest value that is assigned to C is its cluster similarity.
2) Then the Davies-Bouldin index can be obtained by averaging all the cluster similarities.

The smaller the index, the better the clustering result. By minimizing this index, clusters become distinct from each other, and therefore achieve the best partition.

Figure 8: *ELMo (Hierarchical Clustering).* Cluster 0 - Indoor-related Memories

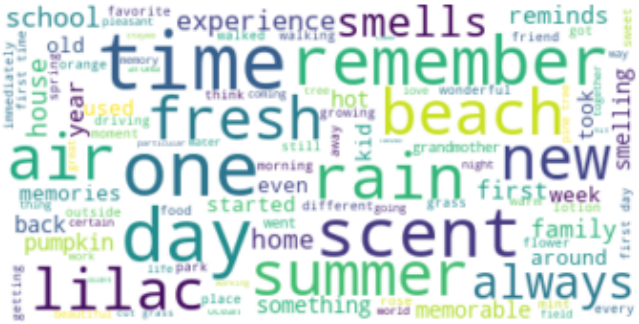

Figure 9: *ELMo (Hierarchical Clustering).* Cluster 1 - Climate Related Smell Experiences

Figure 10: *ELMo (Hierarchical Clustering).* Cluster 2 - Pleasant Experiences

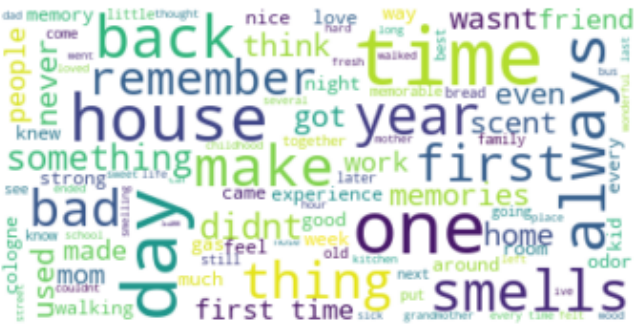

Figure 11: *ELMo (Hierarchical Clustering).* Cluster 4 - Time-related Experiences

Pre-Trained Skip-Gram	
Cluster 0	Negative Emotions - "never, strong, bad, sick, horrible"
Cluster 1	Food Experiences - "coffee, house, bread, pizza, delicious"
Cluster 2	Recalling Smell Linked Memories - "time, experience, day, new, always, used, back, first"
Cluster 3	Pleasant Atmosphere - "scent, perfume, fresh, always, air, sweet"

Figure 12: *Iteration 1*: Pre-trained Skip-Gram Results Showing the Most Popular Words of Each Cluster (Human Analysis)

Pre-Trained CBOW	
Cluster 0	Positive emotions - "first, walked, smells, good, mom, house, gas"
Cluster 1	Pleasant experiences - "perfume, house, scent, always, day, time, fresh"
Cluster 2	Food experiences - "coffee, pizza time, bakery, walking, one, still"
Cluster 3	Undefined - "lilac, kid, cookie, day"

Figure 13: *Iteration 1*: Pre-trained CBOW Results Showing the Most Popular Words of Each Cluster (Human Analysis)

BERT	
Cluster 0	Homely/Comforting Memories - "home, perfume, grandmother, memories"
Cluster 1	Undefined - "first, day, house, time, scent, good"
Cluster 2	Undefined - "time, year, first, one back, day"

Figure 14: *Iteration 1*: BERT Results Showing the Most Popular Words of Each Cluster (Human Analysis)

ELMo	
Cluster 0	Indoor-Related Memories - "remember, new, car, house, room"
Cluster 1	Nature Memories - "remember, beach, lilac, fresh, time"
Cluster 2	Pleasant Experiences - "perfume, memories, scent, cologne, first"
Cluster 3	Food Experiences - "coffee, pizza, morning, food, cookie"
Cluster 4	Unpleasant Experiences - "remember, always, bad, never, wasn't"

Figure 15: *Iteration 2*: ELMo Results Showing the Most Popular Words of Each Cluster (Human Analysis)

SRoBERTa	
Cluster 0	Recalling smell linked memories - "remember, day, food, time, house"
Cluster 1	Food experiences - "coffee, bakery, bread, cookie, remember"
Cluster 2	Pleasant experiences - "perfume, scent, cologne, first, memories"
Cluster 3	Strong Unpleasant experiences - "strong, bad, skunk, used, rotting"

Figure 16: *Iteration 2*: SRoBERTa Results Showing the Most Popular Words of Each Cluster (Human Analysis)

Pre-Trained Skip-Gram	
Cluster 0	Strong Emotional Memories - "sweet, good, bad, memories, strong, first"
Cluster 1	Food Experiences - "delicious, hungry, kitchen, Christmas, Thanksgiving, house"
Cluster 2	Recalling smell linked memories - "old, cologne, house, memorable, open, close"

Figure 17: *Iteration 3*: Skip-Gram Results Showing the Most Popular Words of Each Cluster (Human Analysis)

Pre-Trained CBOW	
Cluster 0	Mixed First Emotions - "house, strong, first, bad, mixed, sweet"
Cluster 1	First Time with Food Experiences - "house, first, sweet, hungry, Christmas, wonderful, good"
Cluster 2	Food Experiences - "kitchen, house, fresh, hawaii, marinara"
Cluster 3	Positive First-time Emotions - "new, best, house, pleasant"

Figure 18: *Iteration 3*: CBOW Results Showing the Most Popular Words of Each Cluster (Human Analysis)

Sentiment Vectors	
Cluster 0	Positive emotions - "good, happy, house, time"
Cluster 1	First time experiences - "first, time, day, always, time, house"
Cluster 2	Negative emotions - "bad, never, something, wasn't"

Figure 19: *Iteration 4*: Sentiment-based Vectors Results Showing the Most Popular Words of Each Cluster (Human Analysis)

	Word Vectors Iteration 1			Sentence Vectors Iteration 2		Enhanced Pre-Processing Iteration 3		Sentiment Vectors Iteration 4
	Pret-Trained Skip-Gram	Pre-Trained CBOW	BERT	ELMo	SRoBERTa	Pre-Trained Skip-Gram	Pre-Trained CBOW	Sentiment Vectors
Silhouette Score (HCA)	0.05375924993564586	0.20990757542109212	0.03317016696606063	0.027744866047784925	0.04601436384157777	0.08838011849377306	0.04760267945370804	**0.5260190667782916**
Silhouette Score (K-Means)	0.0744076824109475	0.1759745787634266	0.07118386669336282	0.04321981351926451	0.03936256152839114	0.10063842598023912	0.07493096914600629	**0.5651390541214006**
Calinski-Harabasz Index (HCA)	31.519201660100865	93.74478086331024	23.122471653907418	13.23165303256513	18.05221472114622	46.426289190296444	25.600450691107064	**446.7319507393843**
Calinski-Harabasz Index (K-Means)	31.75663641301576	104.87247482881173	31.56576787007823	17.225387303454728	42.71602649051107	53.25844442700825	29.925507802317597	**542.5768797815149**
Davies-Bouldin Index (HCA)	3.2610618556289466	2.070169768141664	3.2335632827404637	4.071037778473899	3.9771342343548497	2.843691897926696	2.9539188377782595	**0.7422167967641617**
Davies-Bouldin Index (K-Means)	3.1405329170540854	2.0525339373145655	2.732369143496351	3.4402466624679433	3.8201306268061845	2.8591087247062688	2.7781063915652555	**0.7295641187605263**

Figure 20: *Overview of the Silhouette Score, Calinski-Harabas Index and Davies-Bouldin Index across the 4 iterations.* This version contains more information.

	SAD	HAPPY	SURPRISE	FEAR	ANGRY	DISGUST
Cluster 0	0.00123457	**0.41882099**	0.0617118	0.0018843	0	0
Cluster 1	0.01609484	0.05398742	0.03104949	0.01295883	0.01806133	0.01277242
Cluster 2	**0.12429936**	0.00251823	0.04482888	**0.15979597**	**0.13506476**	**0.1486804**

Figure 21: *Iteration 4: Average Emotions for Each Category Generated.* The bold values highlight the emotions that dominate for each cluster. This version contains more information.