AN NLP APPROACH TO CLASSIFY SMELL EXPERIENCES

A PREPRINT

To Eun Kim*

Marianna Obrist†

Department of Computer Science
University College London
London, WC1E 6BT

January 13, 2021

ABSTRACT

Interdisciplinary studies between human sensory perception and computer science are drawing attention and so are olfaction research, although smell is the least understood among all senses. Ranging from trials to predict natural language descriptions from odour molecules to efforts to classify and model olfactory experiences many attempts have been made to understand the connections between odorant structure, human perception and their experiences from them. However, as research on smell experiences are clearly an area that is under-researched as compared to that on chemical aspect of smell perception, we review recent literatures on olfaction modelling but weighing more focus on smell experiences. Here we identify chasms in research and report the development of self-supervised pre-trained language models, such as GPT and BERT, multi-label text classification, text augmentation and other techniques across the field of Natural Language Processing (NLP) to suggest smell experience researchers to fill the gap and overcome the limitations of current studies by the state-of-the-art NLP methodologies. Application of our suggestion into the research may enable intriguing ideas around olfactory experience to be come true - supporting decision making through smell, smell-enhanced advertisement and fragrance recommendations to name just a few.

Keywords Smell · smell experiences · HCI · multi-label text classification · text augmentation · transformer · pre-trained language models · unsupervised/self-supervised learning · fine-tuning · domain-specific

1 Introduction

Multisensory perception is gaining ground in the fields of Human-Computer Interaction (HCI), Human-Food Interaction (HFI), and many more [23]. However, concerning the human senses, while visual and auditory interaction has been spotlighted as a mainstream research, the number of studies related to olfactory senses and its interaction with human are relatively small. Moreover, little is known about the relationship between olfactory percepts and the language that describe them[9].

Studies on smell can be categorised by the area they are dealing with. Research on relationship between human perception and molecular structure of odorants can be said that they tackle upon chemical and perceptual space of smell. Recent successful trial to predict the relationship between a molecule's structure and odour description [21] falls under the research on chemical and perceptual space of smell. On the other hand, those that deal with experiences and stories of smell can be said that they tackle experiential space of smell. Involvement of personal memories and experiences into this research [17] makes study fall into the area of experiential space of smell.

*to.kim.17@ucl.ac.uk

†m.obrist@ucl.ac.uk | COMP0030 Research Supervisor | Professor of Multisensory Interfaces at UCLIC

Here, in this review, we present an overview of previous efforts to link and model chemical, perceptual and experiential space of smell with more focus on works that deal with experiential space. This is because research on chemical space of smell is relatively more developed compared to that on experiential space. Literature review that can fill the gap between research and can give suggestions to future studies will contribute to the development of this domain. Moreover, smell experiences are central to many lived experiences, because of the role played by odours in people's daily lives - how smells appeal to emotions, the memories they incite and the responses they ignite. This means that there is potential for interesting innovations built around olfactory experiences, which makes this work important and relevant for the future of immersive computing as Obrist et al. addressed[17]. We hereby suggest that current research gaps and limitations can be overcome by approaching the problems with state-of-the-art Natural Language Processing techniques.

2 Related Work

2.1 Chemical space: DREAM Olfaction Prediction Challenge and olfactory psychophysical dataset

One monumental moment in the area of odour prediction is when DREAM Olfaction Prediction challenge[11] is organised in 2015. The goal of this challenge is to form models that can predict characteristics of smell from a molecule's chemical structure. To create a dataset, Keller and Vosshall performed extensive smell-testing to 49 individuals who were required to sniff 476 odour molecules and tell various characteristics of given smell and decide how well the percept aligns with a list of 19 natural language descriptors. With this dataset, participants of this challenge were to predict odour intensity, odour valence and the 19 descriptors that each of 49 subjects have said. Results have shown that regularized linear models and random forest performed well in prediction[11]. All in all, this has opened the gate of this field of research and become a benchmark dataset, thus, still being used by many studies until now.

2.2 Chemical space: State-of-the-art machine learning approach on scent molecules

In 2019, Sanchez-Lengeling et al.[21] have announced record-breaking result for the problem called quantitative structure-odour relationship (QSOR). They proposed the use of Graph Neural Networks (GNN) as a solution to this task. This approach outperformed previous attempts on a dataset labelled by olfactory experts. Additionally, the paper has showed that the stored embeddings retrieved from the learned GNN model can even perform well in downstream tasks by generalising to other olfaction tasks such as the DREAM Olfaction Prediction challenge. This property to perform a transfer learning on chemistry data is a rare ability in current research [21]. Figure 1 depicts an overview how GNN model is structured .

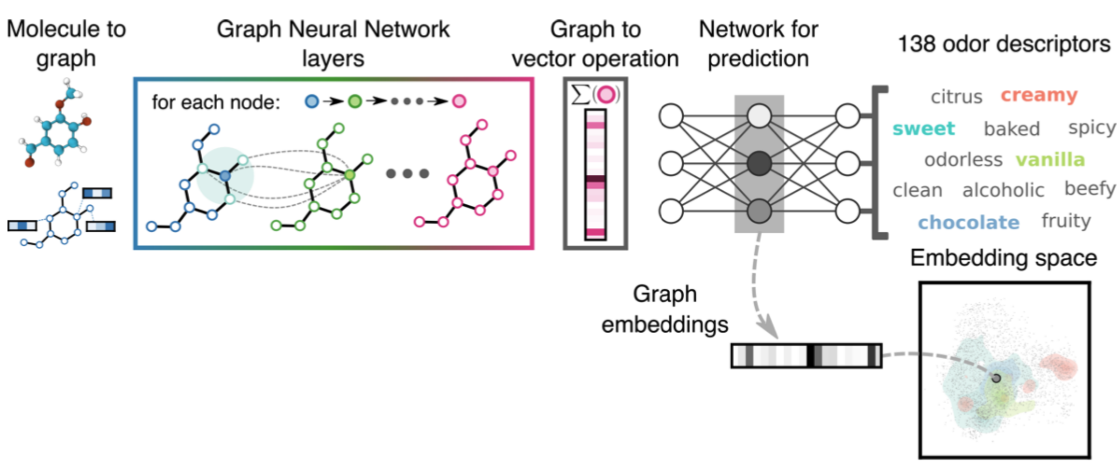

Figure 1: GNN Model Schematic. [21]

A core contribution of this research is that they have successfully built a generalisable model where its knowledge can be transferred into other tasks that tackle olfactory molecules. Useful odour embeddings will be able to perform transfer learning to previously-unseen odour descriptors where no prior methodologies would be able to mimic. Thereby we believe that many future works should and will adopt this model to a number of applications that GNN can bring us.

2.3 Experiential space

As opposed to the chemical space, this branch of research tries to map human's odour perception with individual's experiences or stories related to a certain smell. Research on this area is normally conducted in HCI and is used in technologies that capture and generate smell. Since smell can play a crucial role for our memories and emotions and due to the fact that this is very subjective, we can only link smell blurrily to specific experiences and emotions, which is why we need an extensive survey to coin a dataset [17]. They have collected 439 smell stories, containing descriptions of personal memorable experiences involving smell by distributing a questionnaire though crowd-sourcing. 6 key sections of the questionnaire were about smell story, context, the smell itself, experienced emotions, their view on smell technologies, and personal background including socio-demographic and cultural background. These were categorized into 10 parts and 36 sub-categories then labelled [17].

Their contribution is valuable as they proposed this according to the needs from the academic community in a right time, but they point out some of their limitations: 1. geographical limitation (limited to US), 2. collecting narratives via online lacks sophisticated implications which they could have gained from interviews [17].

Another recent case study by Brate, Groth and Erp[2] on detecting smell experiences in novels is notable as they are one of a very few recent studies on identifying smell experiences from document. Although this study is not about olfactory information classification but about an olfactory information 'extraction', it is still notable as they used semi-supervised approach on narrative data to identify smell from text. They have also created a Literary Smell Dataset which is a manually labelled sentences from English literary. The way the labelled the text and their extraction approaches are new. They syntactically tagged the words by parts of speech and they target certain set of tags, such as adjective and noun groups, to extract olfactory information from novels. The results have shown that the new syntactical approach can offer significantly better performance than the legacy keyword-based model [2].

Still, these have a few limitations. Both datasets by Obrist et al.[17] and Brate et al.[2], have been labelled manually. Similarity-based text clustering [14] method can reduce the burden of painful manual clustering, but this will not be covered further in detail.

3 Filling in gaps

As mentioned above, research on experiential space of smell is immature relative to that on chemical space, which can mean that there are less gaps that can be found in the zone of chemistry. Therefore, in this section, we look for gaps or limitations of smell experience research and find possible solutions in the field of NLP. That is because data they use and collect in current research on smell experience mostly consist of human language data. As NLP has been able to provide answers in wide range of studies so far with great academic histories, we believe that current gaps and obstacles are not different from the tasks that NLP has given answer to, thus can be overcome with similar NLP-based approach.

One of the problems in smell experience research is classification of different smell stories. Dataset created by Obrist et al.[17] has a feature that contains individual's smell stories along with multiple labels as another feature. If we want to classify and identify each smell story, then this comes down to a classical multi-label text classification problem.

3.1 Smell language classification and domain adaptation

Pre-trained Language Models (PTM) based on Transformer architecture [22] has been proven to perform the best on many downstream tasks in the field of NLP. Starting from GPT [19], many variations are keep renewing all-time high record in various metrics: BERT [6], XLNet [27], RoBERTa [15] and GPT-2 [20] are the variations. These models that are trained on multiple tasks in unsupervised/self-supervised way are now becoming a gold standard in NLP research. Therefore, it is natural to suggest employing unsupervised pre-trained language models for doing multi-label text classification on the smell experience data.

There are several reasons why PTMs can be suitable for smell stories classification.
First, pre-training on large text corpus can learn universal language representations and help with the downstream tasks including multi-label text classification.
Second, pre-training can be thought as a type of a regularization to avoid overfitting on small data (suggestions for overcoming small dataset will be dealt more in depth in section **3.2**.
Last but not least, we can fine-tune the model by tweaking the last few layers to train on multi-label text classification task which is also called as transfer learning. Since pre-training provides a better model initialisation, this leads to a

better generalisation performance and fast convergence on the targeted task [7, 6].

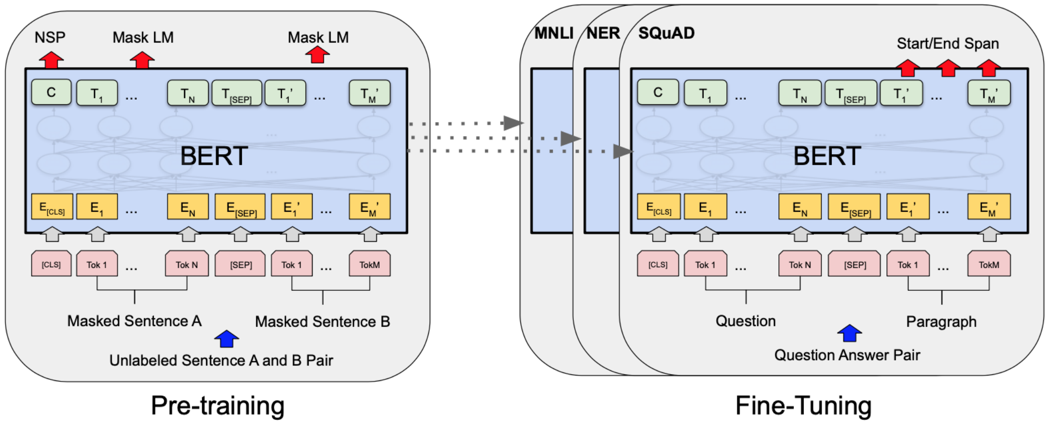

Figure 2: Overall pre-training and fine-tuning procedures for BERT [6]

Figure 2 concisely describes how pre-training and fine-tuning work in BERT. In the phase of pre-training, model does two tasks with tokenized input sentences: masked token prediction and next sentence prediction. Predicting a masked token is shortly termed as Mask LM and this is the nature of self-supervised learning in that it masks some percentage of input tokens by itself and learns the sentence by predicting the most plausible words to be replaced with the masks. Based on the embeddings and weights they have learned, supervised training can be done on a specific task with a labelled dataset as transfer learning [6].

3.1.1 Fine-tune pre-trained models for Multi-label text classification

Out of many examples, this section will address two representative cases of fine-tuned multi-label text classification model: PatentBERT [13] and a BERT-based model trained on EU Legislation [3]. These are the examples that are adapted to the domain of patent-related and EU Legislation language, respectively. Both follow the fine-tuning procedure showed in the original BERT paper, with a little change. They used sigmoid cross entropy with logits function in replace of softmax function, so that they can produce a probability per label. They showed that both of the models outperformed previous best model in their domain[13, 3].

3.1.2 Domain specific pre-training on smell experience data

As mentioned above, domain adaption by transfer learning is being proven to be a fairly robust approach, which can possibly work well on smell experience dataset as well. However, Gu et al.[8], with his team at Microsoft Research, have announced PubMedBERT, which is trained in a way that against a common assumption: out-domain text is still helpful in that it can give general understanding of words (mixed domain pretraining). However, the researchers have seen significant improvement while not using any general texts unrelated to biomedical terms and proceeded pretraining with texts from PubMed only. This is a Domain-specific pretraining, which does not align with the prevailing assumption.

Will this method be likely to work in smell experience data? We presume that it is less likely to perform well with this way. According to the Microsoft researchers, the mixed domain pretraining makes more sense if the target application domain has little text of its own and can thereby benefit from pretraining using related domains. This was not the case for biomedicine, which has around 30M papers in PubMed. Moreover, due to the fact that general domain text is very different from biomedical text, mixed domain pretraining could have rather hindered the target performance in PubMedBERT[8].
However, we know that text that describes one's experience with smell is not very different from general text. Therefore, future researchers had better not to pre-train in domain-specific way on smell experience data even when they are ready with sufficient amount of data.

3.2 Not Enough Data

Dataset of smell story that Obrist and co-workers [17] have built has 439 entries with 10 categories and 36 sub-categories of smell experience. Also, Brate et al.[2] has said that smell experience data accounts for very low proportion in the evaluation set of their literary smell dataset.
Considering that most of the current pre-trained language models must be trained on huge amount of text corpus, it seems impossible to train PTM from scratch with the current dataset – BERT is trained on BooksCorpus (800M words) and Wikipedia (2,500M words) [6]. Although PTM allows us to fine-tune the model with relatively small dataset in a short period of time, we still need more data to achieve high fine-tuning accuracy.

This is where data augmentation can come into play. Data augmentation is used to increase the size of labelled training sets by applying specific transformation on the data while preserving the class labels. It is settled as a common practice in the field of computer vision [12] and speech [5], such as cropping, padding and flipping, but augmentation of textual data is not as common and easy as that of other areas for it distorts the text, making it grammatically and semantically incorrect if we follow the practices in other domains. Thus, careless augmentation of text can do more harm than good [24]. Below presents a number of methods to augment texts with little or no semantical change.

3.2.1 Lexical substitution

This would be the most intuitive way to make variations on texts. This branch of work substitutes words in a text with little or no semantical change based on a few different methodologies.

- **Thesaurus-based substitution**
  We randomly choose a word from a given text and substitute the chosen word with its synonym by referencing a thesaurus table. WordNet [16] can be used as a look-up table. Wei and Zou[24] have used this technique in their 'Easy Data Augmentation' paper.

- **Word Embeddings-based substitution**
  We use pre-trained word embeddings like GloVe[18], and choose the nearest word in the word vector space for the replacement. It became more popular from TinyBERT[10], where they used GloVe in their augmentation procedure to find most similar words of chosen word.

- **Masked Language Model (MLM)-based substitution**
  PTMs like BERT[6] goes through a few pre-training tasks; Masked token prediction is one of the tasks. Models are trained with a deep bidirectional representation by masking some percentage of the input tokens at random, and then required to predict those masked tokens; this process is therefore called as self-supervised learning[6].
  This can be used in augmenting the dataset. Compared to Thesaurus-based and Word embeddings approach, this method will output more grammatically and semantically coherent words as PTMs take context into account.

3.2.2 Back-translation

This refers to the procedure of translating a text in language A to another language B and then translating back into A to get a semantically identical but a varied form of the text for data augmentation[25]. This method is far from lexical substitution but close to a document paraphrasing as can be seen in Figure 3.
This method can be very suitable for multi-label smell experience classification, and it is used in multiple papers and results[25].

3.2.3 Data noising and smoothing

Like we mix some noise pixels in image augmentation to make model bear with perturbations, random noise can also be injected into language data to make a robust language model.

- **Unigram noising for Recurrent Neural Networks (RNN)**
  To directly quote from Xie et al.[26], we define unigram noising as follows:
  "For each x_i in $x < t$, with probability γ replace x_i with a sample from the unigram frequency distribution".

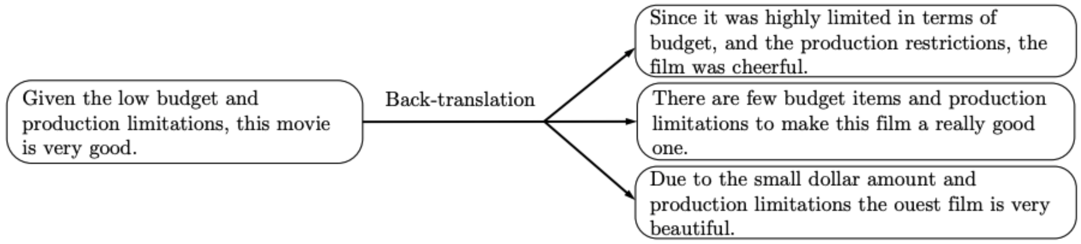

Figure 3: Backtranslation results in text paraphrasing [25]

- **Blank noising for RNN**
  To directly quote from the Xie et al.[26], we define blank noising as follows:
  "For each x_i in $x < t$, with probability γ replace x_i with a placeholder token '_'".

  The idea is to mask random word with placeholder token '_'. The writers use this to avoid overfitting on specific context which helped improve perplexity and BLEU metric [26].

- **Random insertion, swap and deletion**
  Random insertion (we choose a random non-stop word and insert its synonym into a random position), random swap (we swap any two words) and random deletion (randomly remove a word) can improve text classification performance, according to Wei and Zou[24]. The researchers say that these methods can generate a sentence that does not make sense to human but retaining its meaning without overfitting [24].

3.2.4 Syntax-tree transformation

The idea is to parse the sentence and generate a dependency tree to change the sentence from active to passive or vice versa[4].

3.2.5 Language-model-based data augmentation (LAMBADA)

Anaby-Tavor et al.[1] have first come up with LAMBADA for synthesizing labelled data to improve text classification task, and this is especially useful when only a small amount of labelled data is available. Researchers have seen an increased performance compared to the baseline model without LAMBADA. The structure is built upon the robust language model: GPT [19]. The process is as follows:

1. Prepend the class label to each text data
2. With the transformed data, finetune the model (GPT or BERT); fine tuning tasks follow methods from the models' original paper (GPT will generate training data, while BERT will predict masked tokens).
3. After fine-tuning, new training data can be generated once a few initial words including prepended class label are given to the model[19].

4 Discussion and Conclusion

Without a doubt, as many papers have proven, this area of research can pave a way for multisensory HCI and have many use cases on web, advertisements and myriad industries that are directly related or sensitive to smell. Moreover, the founders of the DREAM Olfaction Prediction challenge have mentioned a possibility of reverse-engineering of smell molecules from odour descriptors, which can be a breakthrough in many industrial places[11].

However, especially for studies on experiential space, the biggest problem seems to be the small volume of data present and remaining obstacles in collecting extensive amount of data. Conceiving the moment when DREAM Olfaction Prediction challenge is created, why it is now said to be an historic moment and how it has drawn many people into the research in mind, we should anticipate for more endeavour to create bigger and worldwide dataset with great amount of public engagements. In addition to that, it is also possible to innovate on machine learning frameworks targeted on odorants and olfactory natural language that can provide an high-level API.

Until sufficient volume of smell experience dataset is created, to overcome the issues, we can expect that using pre-trained language models to train a text classifier will produce a fine-tuned embedding that captures the language that represents the classified descriptor, and this may be carried out simultaneously with text augmentations. However, notable is that careful selection of augmentation methodologies is necessary as careless augmentation in NLP can result in misleading outcomes. It is desirable to combine a number of methods to augment the data, because diversity of paraphrases is found to be crucial. For example, one can randomly select which method to use among noise injection, lexical substitution and syntax-tree manipulation, and then feed the output of the transformed text into either back-translation or generative models.

We look forward to seeing an adoption of our suggestion to the research and a growth of smell experience studies in the academic community.

References

[1] Ateret Anaby-Tavor et al. "Not Enough Data? Deep Learning to the Rescue!" In: *CoRR* abs/1911.03118 (2019). arXiv: 1911.03118. URL: http://arxiv.org/abs/1911.03118.

[2] Ryan Brate, Paul Groth, and Marieke van Erp. "Towards Olfactory Information Extraction from Text: A Case Study on Detecting Smell Experiences in Novels". In: *arXiv preprint arXiv:2011.08903* (2020).

[3] Ilias Chalkidis et al. "Large-scale multi-label text classification on eu legislation". In: *arXiv preprint arXiv:1906.02192* (2019).

[4] Claude Coulombe. "Text data augmentation made simple by leveraging NLP cloud APIs". In: *arXiv preprint arXiv:1812.04718* (2018).

[5] X. Cui, V. Goel, and B. Kingsbury. "Data Augmentation for Deep Neural Network Acoustic Modeling". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.9 (2015), pp. 1469–1477. DOI: 10.1109/TASLP.2015.2438544.

[6] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[7] Dumitru Erhan et al. "Why does unsupervised pre-training help deep learning?" In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 201–208.

[8] Yu Gu et al. "Domain-specific language model pretraining for biomedical natural language processing". In: *arXiv preprint arXiv:2007.15779* (2020).

[9] Georgios Iatropoulos et al. "The language of smell: Connecting linguistic and psychophysical properties of odor descriptors". In: *Cognition* 178 (2018), pp. 37–49. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2018.05.007. URL: http://www.sciencedirect.com/science/article/pii/S0010027718301276.

[10] Xiaoqi Jiao et al. "Tinybert: Distilling bert for natural language understanding". In: *arXiv preprint arXiv:1909.10351* (2019).

[11] Andreas Keller et al. "Predicting human olfactory perception from chemical features of odor molecules". In: *Science* 355.6327 (2017), pp. 820–826. ISSN: 0036-8075. DOI: 10.1126/science.aal2014. eprint: https://science.sciencemag.org/content/355/6327/820.full.pdf. URL: https://science.sciencemag.org/content/355/6327/820.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

[13] Jieh-Sheng Lee and Jieh Hsiang. "Patent classification by fine-tuning BERT language model". In: *World Patent Information* 61 (2020), p. 101965.

[14] Y. Lin, J. Jiang, and S. Lee. "A Similarity Measure for Text Classification and Clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 26.7 (2014), pp. 1575–1590. DOI: 10.1109/TKDE.2013.19.

[15] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[16] George A. Miller. "WordNet: A Lexical Database for English". In: *Commun. ACM* 38.11 (Nov. 1995), pp. 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748. URL: https://doi.org/10.1145/219717.219748.

[17] Marianna Obrist, Alexandre N. Tuch, and Kasper Hornbaek. "Opportunities for Odor: Experiences with Smell and Implications for Technology". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 2843–2852. ISBN: 9781450324731. DOI: 10.1145/2556288.2557008. URL: https://doi.org/10.1145/2556288.2557008.

[18] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

[19] Alec Radford et al. *Improving language understanding by generative pre-training*. 2018.

[20] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[21] Benjamin Sanchez-Lengeling et al. *Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules*. 2019. arXiv: 1910.10685 [stat.ML].

[22] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[23] Carlos Velasco et al. "Multisensory Technology for Flavor Augmentation: A Mini Review". In: *Frontiers in Psychology* 9 (2018), p. 26. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.00026. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2018.00026.

[24] Jason W. Wei and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". In: *CoRR* abs/1901.11196 (2019). arXiv: 1901.11196. URL: http://arxiv.org/abs/1901.11196.

[25] Qizhe Xie et al. "Unsupervised data augmentation for consistency training". In: *arXiv preprint arXiv:1904.12848* (2019).

[26] Ziang Xie et al. "Data Noising as Smoothing in Neural Network Language Models". In: *CoRR* abs/1703.02573 (2017). arXiv: 1703.02573. URL: http://arxiv.org/abs/1703.02573.

[27] Zhilin Yang et al. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems*. 2019, pp. 5753–5763.