# Evaluating a relationship between mental health and wearable sensors using NLP

Stanford CS224N Custom Project

**Artem Trotsyuk**
Department of Bioengineering
Department of Computer Science
Stanford University
atrotsyuk@stanford.edu
Mentor: Emma Chen

## Abstract

Here we evaluate 100,000 abstracts published in the last fifteen years to determine if there exist ideal sensors that can more accurately predict onset of stress and depression. We deploy an unsupervised learning approach with information-dense word embeddings to capture complex concepts such as the link between stress and depression. Our initial results indicate accurate word-relationship mappings. Further work will be done to evaluate the dataset.

## 1 Introduction

The majority of scientific literature is published in text format. This results in a lot of research that is released on a daily basis and being able to process this information is getting increasingly complex for humans. Machine learning methods have become more prevalent and accessible for a general user, allowing for implementation of novel methods to process vast amounts of information. One such method is through the deployment of Natural Language Processing (NLP) that allows for an efficient approach to evaluating text, which is what composes most of the research articles published on a daily basis.

In this project, I evaluate abstracts published over the last fifteen years in the fields of mental health, wearable devices and depression to evaluate, in an unbiased manner, if there exist ideal measurement parameters that can more accurately identify onset anxiety and depression. I use an unsupervised learning approach with information-dense word embedding to capture complex concepts such as the underlying link between sweat and onset on a panic attack, using a Word2Vec architecture as done in [1].

## 2 Related Work

Swain et al., authors describe a toolkit that allows for the extraction of specific information from scientific literature by using tokenization and phrase parsing [2]. Their F1 score is near 90% overall, which shows how effective NLP can be in quick data extraction. Similarly, Müller et. al., evaluated gene ontology and compared it to biological literature [3]. Their system, Textpresso, looks at papers, splits the sentences into words/phrases and labels them in a lexicon map. Subsequently they evaluate the lexicon map to documented ontology in the biological science field. Limitations to the above methods, including others

published, include the need for large hand labeled datasets in order for the models to identify appropriate word relationships.

In the work published by Tshitoyan et. al., the authors describe a method that builds on a commonly used word2vec model, which they describe as materials2vec [1]. With this model, the authors evaluated abstracts in the materials science space and determining if there exist more optimal combinations of chemicals on the periodic table for an optimal battery design. The following work subsequently evaluates the model reported by Tshitoyan et. al., and applies it to mental health and wearable sensors.

# 3 Approach

## 3.1 Architectures

For the purposes of beginning evaluating the performance of the model on the dataset obtained for this project, a skip-gram model architecture was deployed, as seen in **Figure 1**. The training objective is to learn word vector representations that can predict nearby words. As a baseline for evaluating the model, I used the skip-gram model architecture, as established in the original Word2Vec study [4].

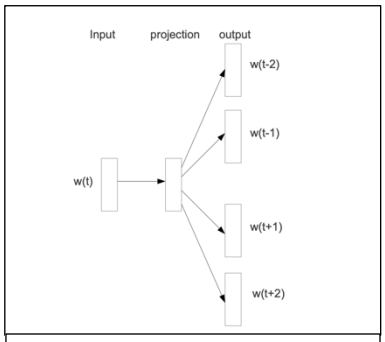


**Figure 1:** Skip-gram model architecture. Figure from original Word2Vec paper [4].

The goal of this project is to evaluate unlabeled language data with a modified Word2Vec model. In this model, outlined in Tshitoyan et. al, each word is associated with a vector. These vectors encode relationships and analogies between words. **Figure 2** illustrates linear vector relationships to better explain word relationships.

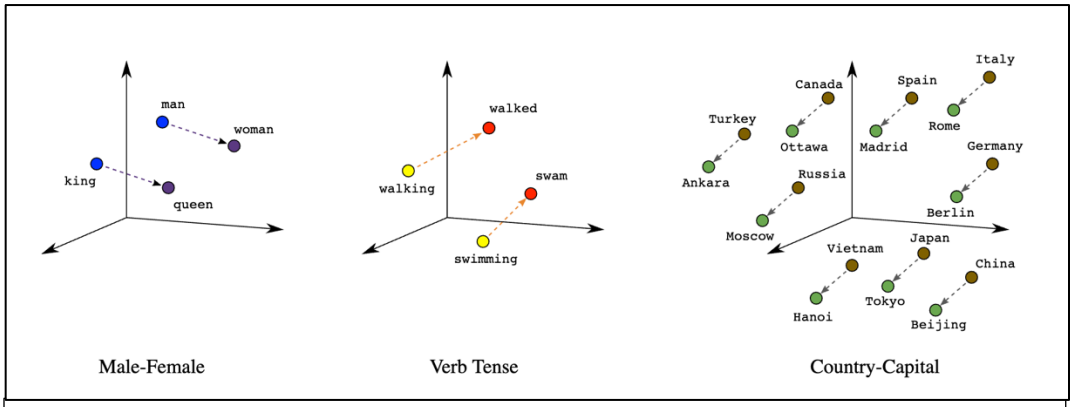


**Figure 2:** Visual explanation of linear vector relationships to explain word relations.

Above, we see stronger relationships between similar analogies such as “king is to queen as man is to woman” or “Italy is to Rome”. Building on a baseline of known analogies and word embeddings, my project sought to find unknown relationships between words. These relationships may help find latent knowledge in research papers. In doing so, I would be able to observe if there are relationships, if any, between sensors and wearable technologies to the ability to predict a mental health state.

# 4 Experiments

## 4.1 Data

My dataset contains 100,000 abstracts pulled from Elsevier's API. I used keywords that include: cortisol sensing; sweat rate; biomolecular sensors; stress; depression; saliva, patch; biosensor; mental health; bioreceptor; proteins; electrical detection; continuous monitoring; cytokines; IL-6; Th-1; Th-2; amygdala; hippocampus; limbic system. The goal of keyword optimization is to mine abstracts for "topic-specific" terms that would subsequently pull relevant abstracts.

## 4.2 Evaluation Method

Initial evaluations deployed to validate the original materials2vec model [1] were to determine if the model was establishing proper similarity distances between words. In the diagram below, we see an example of the central methodology of evaluation used to evaluate the original dataset by Tshitoyan. When the model evaluated "thermoelectric" for example, second order similarities were pulled that contained chemical elements such as $Li_2CuSb$, $Cu_7Te_5$, and $CsAgGa_2Se_4$ **(Figure 3).** These elements were then validated by a baseline, which in their case was first principles understandings of chemical compounds based on the periodic table.

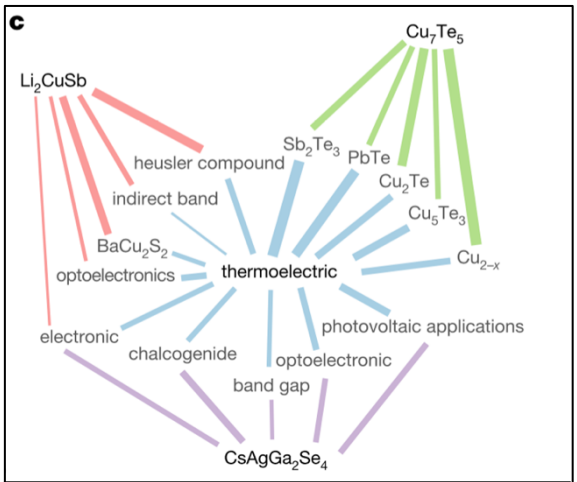


**Figure 3:** Materials2Vec word similarities based on unsupervised abstract evaluation.

Similarly, in order to evaluate the validity of my results and have a true baseline for comparison, I recruited clinicians and engineers to evaluate the output and rank the success of the result based on their domain knowledge on a scale of 1 to 5 (1 being that the model output does not make sense and 5 being that the model output makes complete sense). In doing so, I have a third-party evaluator provide me insight to determine unbiased ground truths that the model can be compared against.

## 4.3 Experimental Details

My project was to replicate the work done in the original paper [1], using a different dataset in the workflow. I went through the original source code, simplifying it in a Jupyter Notebook. I subsequently obtained an API from Elsevier which I included into the updated source code. I then ran a model on 100,000 abstracts and evaluated cosine similarities of keywords of interest. I finally plotted the data results in a reduced dimensionality format, comparing against the baseline Word2Vec model to determine if the similarity clustering was relatively appropriate given the initial model structure and keyword parameters. I fine-tuned model parameters based on known analogies such as "cortisol – stress; saliva – biomarker; patch – sensor; amygdala – depression; etc."

## 4.4 Results

My initial results show sufficient baseline progress in mapping words in similarity space. When I optimized the model further with appropriate analogies and increased my dataset size, I was able to obtain output cosine similarities that were ranked higher by my third-party observes. As an example, initially my model was suggesting that "cortisol-non pregnant" had a similarity score of 1.51. After tuning, the model suggested that "cortisol-females" was more similar, with a similarity score of 1.33 **(Figure 4).**

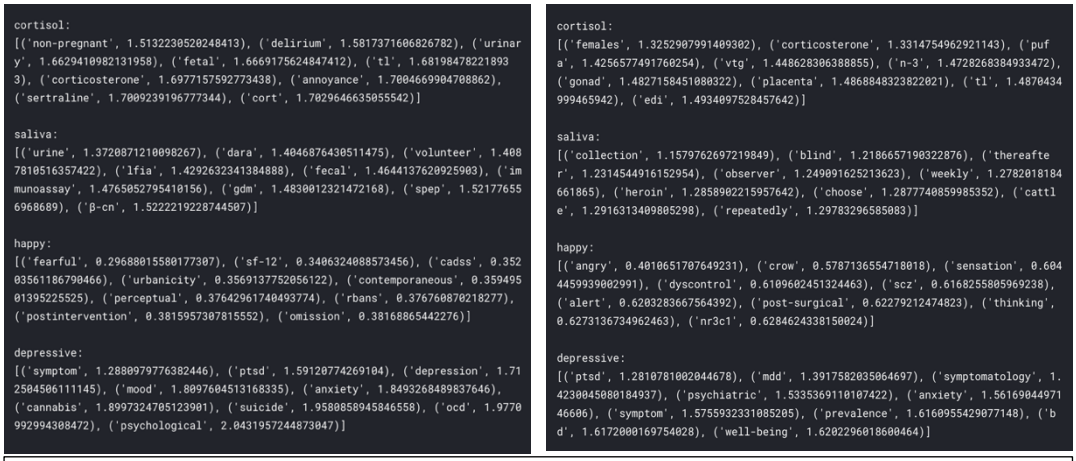


**Figure 4**: Comparison of cosine similarities pre (left) and post (right) tuning of model trained on 100,000 abstracts.

Similarly, when I evaluated the data by reducing the dimensionality into two principle components, I noticed similar clustering of words such as depression-mental-disorder. Interestingly, when the model was tuned, clustering around cortisol shifted. We see words such as amygdala and receptor have reduced distance as compared to pre tuning where we observed cortisol being close to major and brain **(Figure 5).**

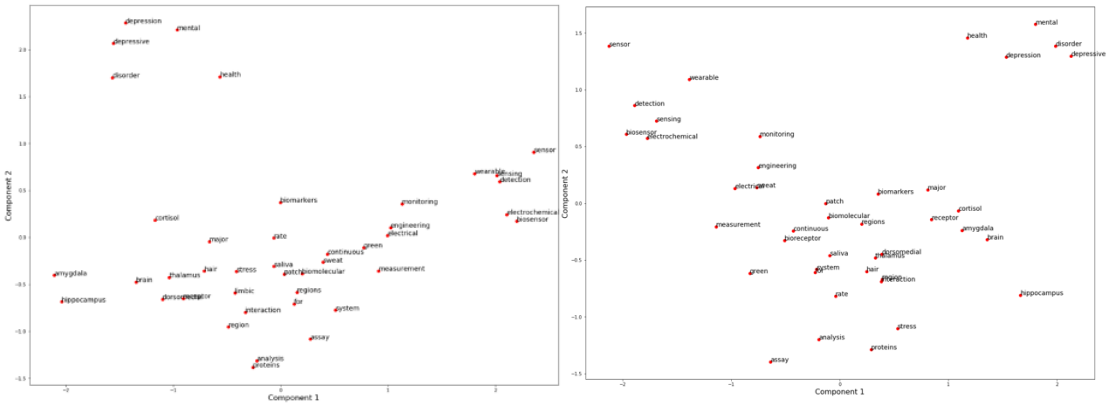


**Figure 5**: Comparison of output data reduced into two principle components, pre (left) and post (right) tuning of model trained on 100,000 abstracts.

To validate the results of the tuned model, I consulted with clinicians and engineers in order to determine whether the tuned model enabled for more accurate similarity prediction. A total of 3 individuals per category were chosen. Clinicians included a surgeon, a psychologist and a neurobiologist (all of whom have direct knowledge regarding treatment and/or diagnosis of depression). Engineers included a chemical engineer, an electrical engineer and a materials science engineer (all of which have previous or current knowledge in wearable sensor technologies). Each individual was a blinded observer. The only context given was the scope of the work and the task I wanted them to complete – rank the clustering results based on domain expertise on a scale of 1 to 5 (1 being horrible and 5 being relevant/accurate in scope).

**Table 1: Validation by Domain Experts**

| Domain Expert (n = 3) | Cosine Similarity Ranking | PC Clustering |
|---|---|---|
| **Clinicians** | 4 | 4.5 |
| **Engineers** | 3.5 | 4 |

A notable observation included the fact that engineers found the data to be less informative that clinicians. The clustering observed for sensors did not indicate significant novel information. Clinicians on the other hand, viewed the tuned model as much more clinically relevant in the data output. Both groups did believe that the observed clustering was more accurate after parameter tuning. This can indicate that tuning based on analogies enables identification of word similarity cores more accurately. My quantitative results were at a level of expectation that I felt was better than if we were evaluating data at random. Since the approach of Tshitoyan was to evaluate abstracts based on an unsupervised manner, the output data obtained enabled for potential novel evaluation of sensors, wearables and mental health evaluation.

# 5 Analysis

The output that I obtained was interesting when it comes to evaluation of sensors and depression. Because the initial analogies used were contextually specific to keyword that are known to be accurate word analogies, it is difficult to determine whether this model enabled for a true unsupervised approach to evaluate scientific abstracts. What was most interesting was observing the clustering of words based on known similarities. Having depression, mental and disorder cluster together signifies that the model is accurate at determining word similarities. The model should be further improved with more data and more fine tuning of analogy parameters. This way the model may draw better similarity rankings when we look at cortisol and think about how that links to depression.

# 6 Conclusion

This project allowed for me to evaluate an existing model and adapt it to a different dataset. The initial architecture was tuned in order to fit the research question at hand. What I learned from this project was that perhaps an unsupervised approach to evaluating contextual similarities may not be the most optimal method of data mining. In order to be

able to have a more robust evaluation of data in the model I was evaluating, we would need to have far greater amounts of data. This in itself was a challenge as publishers do not enable for large data dumps within their APIs, which creates a limit to how much data can be accessed at a given time. Future follow up work would allow for the configuration of this model with a larger dataset, assuming that Elsevier enables research access at greater amounts, as well as trying other more robust models for evaluation such as FastText, BERT and GloVe. This approach will sufficiently evaluate performance of the updated materials2vec model on a novel space of mental health and wearable devices.

# 7 References


[1] Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 2019.

[2] Swain and Cole. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. J. Chem. Inf. Model., 2016.

[3] Müller et al. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLOS Biology, 2004.

[4] Mikolov, T. et al. Distributed representations of words and phrases and their compositionality. Preprint at https://arxiv.org/abs/1310.4546 (2013).