Self-Love in the Context of Alcohol + Drug Recovery

Project Advisor:
Kelly L Ziemer
PhD Candidate in Social Welfare

Team Members:
Anika Cruz
Genevieve Li
Jianing (Jenna) Yu

The Problem

Our project attempts to answer these main questions:

- How is self-love discussed on social media by people referencing alcohol and drug recovery?
- Some examples of how they may discuss self-love could be:
- What differences are there in how #selflove is discussed in a general context compared to AOD recovery?
- To what extent can we predict AOD recovery content within #selflove posts?

The purpose of our project is to examine, through machine learning and a content analysis, the phenomena of #selflove on Instagram and Twitter to see how it is discussed within the context of AOD recovery.

The Data

Two data sources were used in our work: Instagram and Twitter.

The Instagram data is generated from Instagram using Instaloader, a tool to download pictures, videos, captions, and other metadata from Instagram in JSON, txt, jpg, etc. When we collected the data, we set the parameter to collect posts made in 2019 and tagged #selflove.

Some of the biases or problems within this dataset that we found were
(1) advertisements posts
(2) plagiarism posts
(3) no meaningful content posts (e.g., just images, just hashtags)
(4) non-English posts

To solve this kind of issue, we proceeded to eliminate those posts with the preprocessing.

Preprocessing

Our steps to clean and preprocess the datasets include the following:

Lowercased all the words in our dataset.

Removed usernames utilizing Regex.

Made sure that all contractions were **decontracted** to ensure that words like “haven’t” and “have not” are treated the same way.

We **added spaces** to illogically combined words.
This involved finding a solution to an issue encountered during our first iteration of preprocessing and with the Python package wordninja. Our solution involved training a custom language model to identify words that should not be split up. Now, following punctuation removal, words like “self-love” are corrected to “selflove.”

Removed Punctuation from our dataset by utilizing Regex.

Removed Digits from our dataset using a list comprehension to check whether or not words in each document of the corpus is not a digit.

Reduced Lengthened Words. For instance, words such as “loveeeeee” are corrected to “love."

Applied **tokenization** to our data to obtain the token from each word.

Spell Checked our dataset in order to correct any misspelled words.

Removed Stopwords such as “for”, “the”, etc. from our dataset. We also added more stopwords to the existing corpus and to remove from our data.

Ran lemmatization to obtain the stems of words in our data. Originally, we used stemming to achieves this but we encountered an issue in our first iteration of preprocessing. Many words in our data were incorrectly spelled or incomprehensible. Words such as “pron”, “co”, “happi”, “raf”, “los”, “beauti”, “uk”, “ed”, and “anyth” were common words we saw. As such, we ended up choosing lemmatization as a better alternative to stemming.

Exploratory Data Analysis

EDA on Full Dataset
15,773 posts are from IG
175,999 posts are from Twitter
191,772 posts in total

1,527,637 hashtags used
138,367 unique hashtags
Most commonly used hashtags: #selflove' ‘#selfcare', '#love', ‘#motivation’, ‘#inspiration'

85,273 unique user IDs from our dataset
Average number of posts per user id is 2.25

EDA on AOD vs. non-AOD Dataset
892 are AOD posts
959 are non-AOD posts
1851 posts in total

23548 hashtags used
4380 unique hashtags

Most commonly used hashtags: #selflove', '#recovery', '#selfcare', '#soberlife', #sober','#motivation', '#sobriety'.

1183 unique user IDs from our dataset
Average number of posts per user id is 1.57

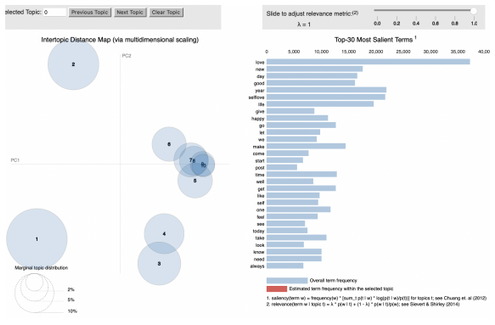
Modeling

Topic Modeling

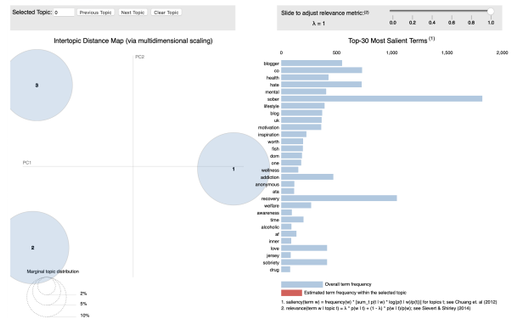
In order to determine and investigate the main topics of our full dataset and our AOD vs. non-AOD dataset, we used different clustering techniques: K Means clustering using LDA and using pre-trained word embeddings. We pursued different clustering techniques to compare the resulting clusters and determine which outputs were more interpretable. Below are examples of our clustering results.

K Means Clustering using LDA

FULL

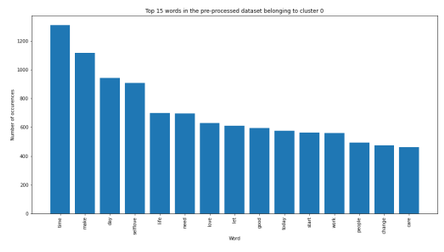


AOD

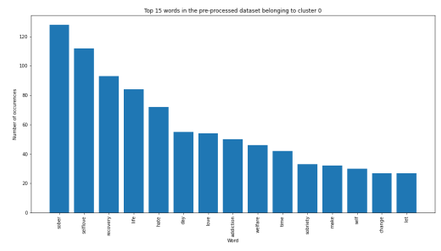


K Means Clustering using Word2vec embeddings

FULL

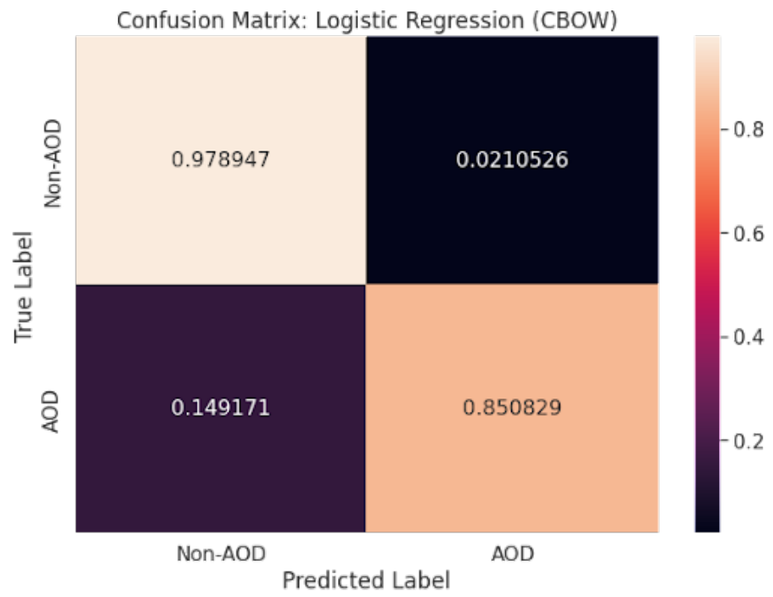


AOD



Classifying AOD vs. non-AOD

To answer our research question about detecting AOD recovery content in social media posts, we trained a classifier. We wanted to be able to classify AOD vs. non-AOD posts in the most accurate way possible. In order to do so, we had to perform feature extraction on the preprocessed AOD vs. non-AOD data. We decided to try three different types of word embeddings: term frequency-inverse document frequency (TF-IDF) and Word2Vec: Skip-Gram (SG) and Continuous Bag of Words (CBOW) in order to represent our words mathematically.



Furthermore, we utilized four different models: logistic regression, naive bayes, XGBoost, and decision tree. The combination that obtained the best classfication results ended up being the continuous bag of words feature matrix using a logistic regression model.

Insights

First, the results from K Means clustering using LDA on our full preprocessed dataset display 10 clusters and the top 30 most relevant terms for each cluster. The results of our hyperparameter tuning found that the optimal number of clusters (k) was 10 clusters. This suggests that our full dataset covers a range of topics. Next, the results from K Means clustering using LDA on our AOD dataset display 3 clusters and the top 30 most relevant terms for each cluster. The results of our hyperparameter tuning found that the optimal number of clusters (k) found was 3. This suggests that this data covers a smaller range of topics. But, since time did not permit our team to interpret these clustering results using domain knowledge, our next steps for topic modeling will involve and depend on our annotation team’s interpretation of the clusters from our full data and our AOD data.

The results from classification for AOD vs. non-AOD posts was quite successful using a continuous bag of words feature matrix along with logistic regression. We were able to achieve an accuracy of ~0.91 with the classifier. Additionally, we plotted the DecisionTree for the cbow and skip-gram feature matrices, respectively. We are limited in the amount of insight gained from plotting the two DecisionTree models since we are currently unable to ascertain the feature labels. Therefore, additional actions need to be taken for this part of the modeling process.