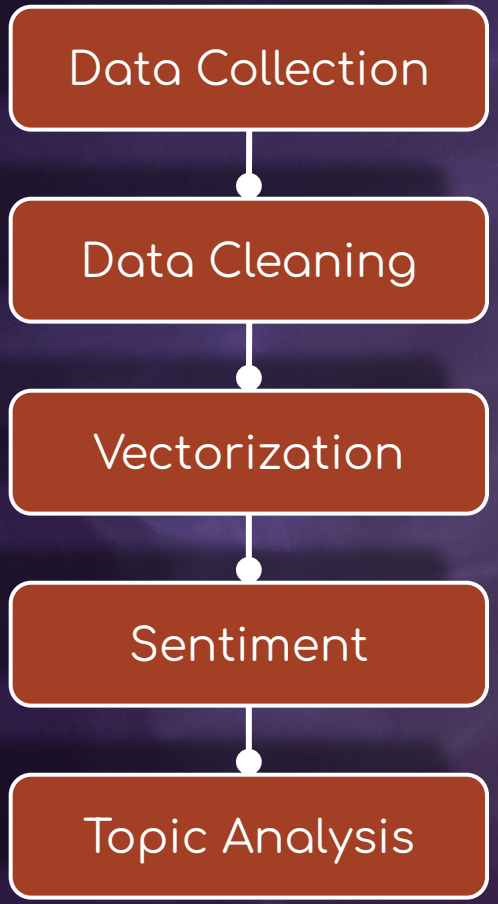
Text Mining to Understand E-Cigarette Usage

Angelo Rosso

Workflow Overview



Data Collection

Research

Browsed reddit for subreddits related to e-cigarette usage

Selection

I created an unbiased list of subreddits expected to show many viewpoints

Quantity

I collected 1000 posts from each and all of the first two levels of comments

Subreddits used: electronic_cigarette, VapeWild, QuitVaping, Vape_Chat, juul, E_Cigarette

Data Cleaning

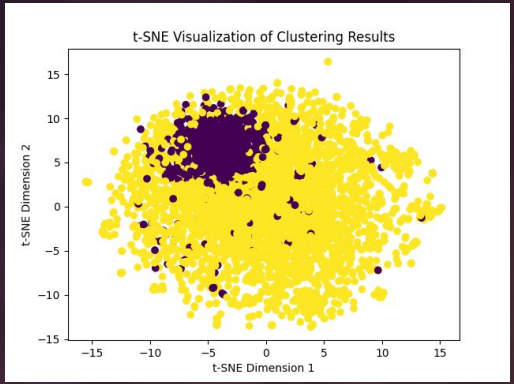
I haven't smoked in 5 years thanks to https://smokefree.gov and I have never been better!

I smoke year thanks URL I never better

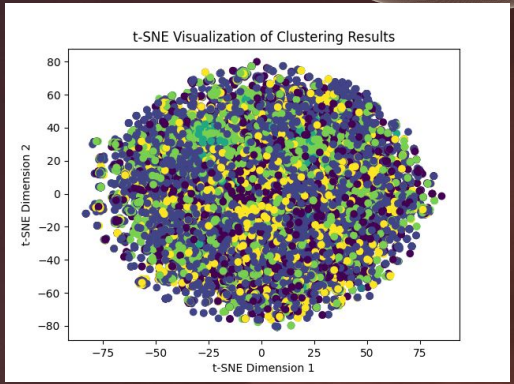
- Filtered out irrelevant comments using custom wordbank
- Generalized all URLs, usernames, and subreddit references
- Tokenized
- Lemmatized and removed stop words
- Convert contractions to words
- Removed special characters, numbers, and punctuation
- Removed too short comments

Word filter: vape, vaping, smoking, smoke, ecig, ecigarette, liquidnicotine, electronic, cigarette, e-juice, e-liquid, ejuice, eliquid, ehookah

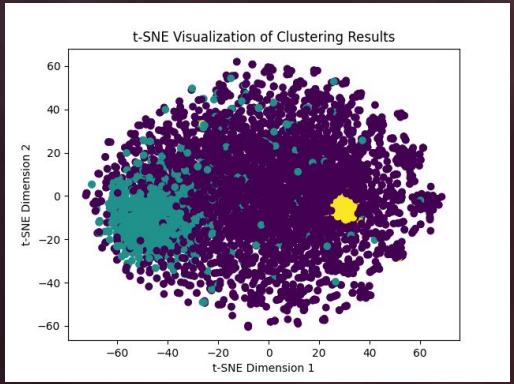
K-Means Clustering (T-SNE visualization)



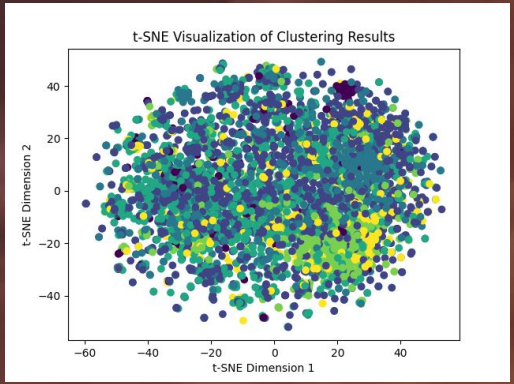
Clustering of Minecraft and Electronic Cigarette data



Clustering of all vaping data



Clustering of Minecraft, Electronic Cigarette, and QuitVaping data



Clustering of smaller sample of vaping data after sentiment analysis

Sentiment Analysis

Vectorized using TF-IDF vectorization and analyzed sentiment using VADER

Positive

Comments with sentiment higher than 0.1

Number of positive: 15197

Neutral

Comments with sentiment between -0.1 and 0.1

Number of neutral: 4568

Negative

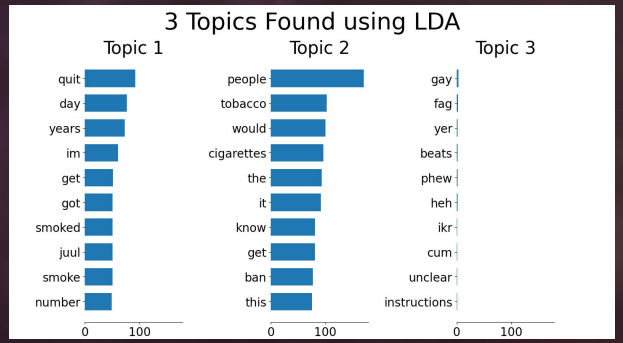
Comments with sentiment lower than -0.1

Number of negative: 8057

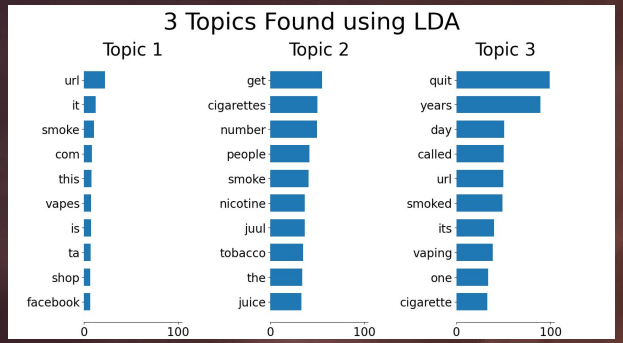
Average sentiment: 0.183

Topic Modeling

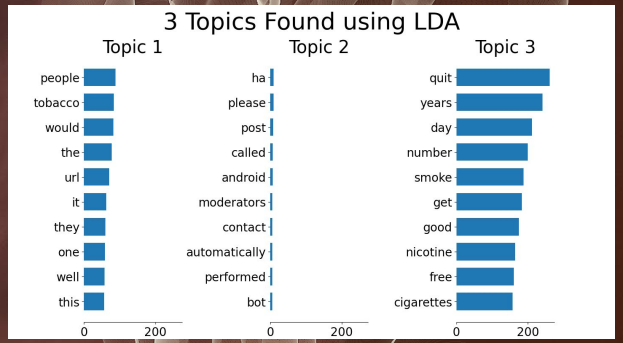
Analyzed topics using LDA; best results after sentiment division



Negative Topics



Neutral Topics



Positive Topics

Sentiment Results

	Positive Post	Neutral Post	Negative Post	Row Sum
Positive Discussion	794	1048	342	2184
Neutral Discussion	810	1298	358	2466
Negative Discussion	124	232	150	506
Column Sum	1728	2578	850	5156

Per Comment

Average sentiment: 0.183

29% | 16% | 55%

Per Discussion

Average sentiment: 0.169

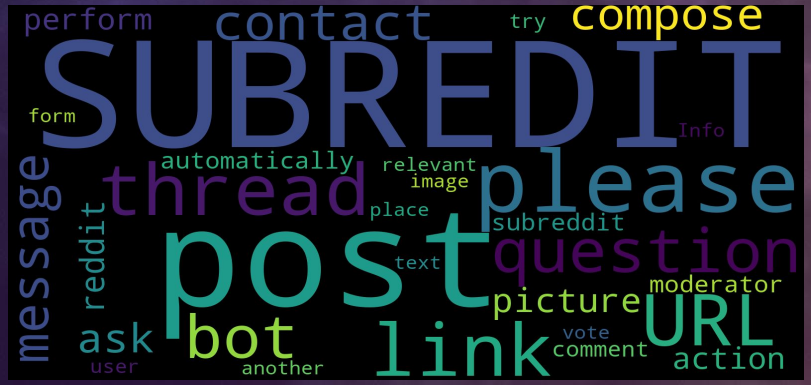
10% | 48% | 42%

Sentiment Results (cont.)

Positive discussions were 6.4 times more extreme than negative discussions

Average negative discussion sentiment: -0.0314
Average neutral discussion sentiment: 0.0002
Average positive discussion sentiment: 0.2008

Topics Results



An example of the "administrative" topic

An example of the "testimonial" topic

- 3 topics worked best for each sentiment
- "Testimonials" are most common topic in **all** sentiments
- Negative testimonials focused on negatives of tobacco/cigarettes
- Negative topics: testimonials, trolls
- Neutral topics: testimonials, acquiring paraphernalia
- Positive topics: testimonials, administrative

Conclusion

Key Findings

- Most users seems to be switching from smoking cigarettes
- Most engagement is positive even in subreddits I expected to be negative like quit_vaping
- The comments that are negative mainly focus on tobacco and cigarettes
- Personal stories are the most common topic of engagement

Focus on positive effects of stopping, not negative effects of continuing

Limitations

Clustering

- Even with hyperparameter search it's hard to choose how many clusters
- Intra-cluster distance stayed constant and inter-cluster distance rose with number of clusters
- Can cluster based on single words

```
Dude looks amazing keep going awsome|*|1
actually fucking coolest thing ever please continue adding stu
see front page, guy|*|0
I front page See|*|0
Looks dope Keep going|*|1
Please keep going|*|1
I think creeper left alone unnerving emotionlessly pass like i
Honestly I think angry face better, thought creeper coming spe
Dude thats awesome keep going skeletons next|*|1
But since solid bone animations could possibly|*|0
Keep going looks amazing. I want see others make|*|1
```

Sentiment Analysis

- Negatively classified comments are not necessarily negative about vaping

```
Bad breath. No taste. Can't breath. Fuck smoking. -> -0.431
```

- Sentiment is very dependent on processing

```
I quit vaping. I don't feel like shit anymore -> 0.2057
I quit vaping I feel shit anymore -> -0.5574
```

Filtering

- I filtered out many comments that were talking about vaping and just didn't have any words in the filter list