Text Mining to Understand E-Cigarette Usage

SPC TP Intro To Data Science (Summer 2023)

Due: June 30th @ 11:59 pm

Motivation:

Social media (SM) platforms such as Twitter and Reddit are rapidly becoming key resources for public health studies. Vast amounts of freely available, user-generated online content allow for studies of public sentiment and informedness on a variety of health-related issues, such as smoking and vaccination, as well as the discovery of emergent patterns that may not be readily detectable using traditional surveillance methodologies, including pre-formulated surveys. If identified and surveilled, SM data can reveal the perceptions, sentiments, and the overall informedness of the general population regarding health problems that are affected by their choices and behaviors, providing opportunities to remedy any misperceptions and improve health outcomes. However, the sheer volume of SM data makes it impossible to organize and retrieve the relevant information in a straightforward manner.

Natural language processing (NLP) provides a wealth of techniques for studying large volumes of text. However, most studies have relied on manual analysis for sentiments, content, and networks. This reliance on manually analysis presents some problems:

1. Due to the time and labor-intensive nature of manual analysis, only a small number of tweets or posts can be analyzed;
2. Due to limitations on how much data can be annotated, those annotated tweets may not be representative of the complete phenomenon.

In your project, you will use data-mining methods to analyze large amounts of Reddit data to automatically gain insight into the topics and sentiments discussed with respect to e-cigarette usage. These insights will help policy-makers to understand the reasons, rewards, and barriers which drive behaviors such as adopting, continuing, or stopping e-cigarette usage.

Limitations of Previous Study:

This is a follow-up study on an initial exploration of the problem. In the initial study, Twitter data was analyzed, however there were several limitations that you will address:

1. We found that Twitter data isn't very informative
    a. Tweets tend to be short and lack information. The content of many Tweets were images, short quips, brags, advertisements or reposts.
    b. We believe Reddit is a better source of relevant text information. Since users tend to discuss issues
2. We didn't use any sentiment analysis tools
    a. The clustering and topic modeling applied didn't fully capture the meaning of data. Sentiment is an important aspect of speech which was not modeled. For example, a topic may be focused on e-cigarette smoke which may be perceived positively by users but negatively by non-users.
    b. We believe sentiment analysis tools should be used within the data mining pipeline
3. We didn't specifically filter out marijuana Tweets
    a. Many Tweets discussed marijuana-vaping, rather than nicotine. This study is specifically interested in understanding nicotine-vaping.
    b. We should be more selective in data we collect or better filter out marijuana-related content

Your Work:

You will perform a follow-up to this previous study and address the limitations found (see point b under each limitation). You will:

1. Data Collection - gather Reddit data
2. Data Cleaning - clean, normalize, and filter the data
3. Data Mining - use clustering, topic modeling, and sentiment analysis to gain insights into the data
4. Iterate over this process to clarify and gain more insights

5. Summarize the insights using visualizations (plots, word clouds, etc), examples, and your own best judgment

Deliverables and Grading:

Submission is via GitLab and Scholar:

1. Via Gitlab: your Python code and data.
    a. You must have at least 3 files which run with a single click:
        i. collect_data.py – this collects data from Reddit and saves it to file
        ii. process_data.py – this cleans and filters data and saves it to file
        iii. analyze.py – this vectorizes, analyzes, and visualizes the data
    b. You may have other supporting files, but these are the primary three files I will be looking at to verify the correctness of your process.

You will be graded on the completeness and readability of your code. Don't have lots of code commented out. Add your own comments. Remove unused and unnecessary files - this means removing or at least moving my example code to a different folder. Make the code nice so that you can reference it a year from now. You must collect data. You **must** perform data cleaning. You **must** at least try clustering, topic modeling, and sentiment analysis. However, how you use these tools is up to you. You may use all, or some in combination, and you can use any other tools available online. How you combine them requires manual analysis and critical thinking. Lastly, you must do something to summarize your results. Use your best judgement to **create a convincing presentation**.

2. Via Scholar: a video-presentation and slides describing:
    a. Details of your data collection and data cleaning process
        i. How did you collect and filter the data?
    b. Details of your data-mining process
        i. What tools did you use and in what order?
        ii. A flow chart is really helpful in explaining this
    c. An example of an iterative development
        i. You should analyze data and learn some ways to improve the downstream process and repeat. You should do at least once, but multiple times is better. Tell me what you did
    d. The insights you found
        i. You must support these insights with evidence (e.g. examples, statistics, and/or visualizations)

Pretend I am a policy-maker with little technical experience. You will be graded on the strength of the insights you found and how convinced I am that they are true. -- Will I, as the policy maker, be willing to put a million dollars behind an advertising campaign to correct misperceptions or pass/repeal a law using the information you provide?