2019

Automated Content Analysis to Quantify Valence, Themes, and Norms Mentioned in Tobacco and E-cigarette Coverage across Two Social Media Sources

Laura Gibson
gibla@pennmedicine.upenn.edu

Sharon Williams
sharon.williams@asc.upenn.edu

Leeann Siegel
leeann.siegel@asc.upenn.edu

Chae Lee
chaerlee@sas.upenn.edu

Allyson Volinsky
allyson.volinsky@asc.upenn.edu

See next page for additional authors

Follow this and additional works at: https://repository.upenn.edu/asc_papers

Part of the Communication Commons

Recommended Citation

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/asc_papers/688
For more information, please contact repository@pobox.upenn.edu.
Automated Content Analysis to Quantify Valence, Themes, and Norms Mentioned in Tobacco and E-cigarette Coverage across Two Social Media Sources

Disciplines
Communication | Social and Behavioral Sciences

Author(s)
Laura Gibson, Sharon Williams, Leeann Siegel, Chae Lee, Allyson Volinsky, and Robert Hornik
Automated Content Analysis to Quantify Valence, Themes, and Norms Mentioned in Tobacco and E-cigarette Coverage across Two Social Media Sources

L.A. Gibson¹, S.E. Williams², L.N. Siegel², C.R. Lee², A.C. Volinsky², and R.C. Hornik²

This file contains headers that may not be able to be expanded in some versions of Word (e.g., for Word 2016 for Mac). If unable to expand headers, please refer to the version of this document with the headers already expanded.

Please note: The social media examples in this document were randomly selected and their content may be offensive, explicit and/or not safe for work.

This research was supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) and FDA Center for Tobacco Products (CTP) under Award Number P50CA179546 (the University of Pennsylvania Tobacco Center of Regulatory Science (UPenn TCORS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Food and Drug Administration (FDA).

¹ Medical Ethics and Health Policy, Perelman School of Medicine at the University of Pennsylvania
² Annenberg School for Communication, University of Pennsylvania
Automated Content Analysis to Quantify Valence, Themes, and Norms Mentioned in Tobacco and E-cigarette Coverage across Two Social Media Sources

In this study, we used supervised machine learning and dictionary coding to measure the coverage of valence, tobacco control themes, and individual norms in content related to (1) tobacco and e-cigarette products in (2) social media sources (Twitter and YouTube). Tweets were collected from May 18, 2014 – December 31, 2017; videos were collected from June 30, 2014 – June 30, 2017. The following example tweets/videos were pulled from the 75,322,911 tweets and 12,262 videos collected through June 30, 2017.

Corpus

We first searched the GNIP and YouTube databases for tweets/videos with a list of hundreds of keywords and search rules that included specific product types (e.g., tobacco, hookah, nicotine, e-cigarette, e-juice), names and brands (e.g., Newport, Skoal, Njoy), behaviors (e.g., vape, vaping), and public health/policy terms (e.g., Quitlines, CDC tips, TruthOrange). This was a broad search meant to capture the majority of tobacco-related content. We then used automated coding methods with validation to clean the database more precisely to tobacco-related content. Tweets/videos came from:

1. **Twitter**
   - n=75.3 million tweets, 32% e-cigarettes

2. **YouTube**
   - n=12,262 videos, 75% e-cigarettes
   - Videos were limited to those with at least 10,000 views in the first 180 days after they were posted. The titles, descriptions, tags and captions of popular YouTube videos were used in the classification process.

Tobacco (no ecig) versus E-cigarettes (ecig)

We define e-cigarettes as electronic devices typically meant for nicotine delivery though they can be used with or without nicotine. These are relatively new products relative to other tobacco products. Any tweets/videos which contain a mention of e-cigarettes are coded as “ecig”. Our definition of “tobacco (no ecig)” products includes cigarettes, cigars, cigarillos, little cigars, smokeless tobacco, and hookah – all non-ecig products used without marijuana. It is possible that there are some mentions of these other tobacco products in the tweets/videos which contain e-cigarette mentions. We decided to count both tweets/videos that are only about e-cigarettes and tweets/videos that are about e-cigarettes and other products as “ecig”.

Valence, Themes, and Norms coded

We used supervised machine learning and dictionary coding to measure the coverage of valence, tobacco control themes, and norms in this corpus of tweets/videos. Experts coded samples of ~2,000 tweets/videos which were split into training and test sets for developing automatic algorithms. Reliability was >.80 for themes, >.79 for norms, and >.73 for valence.

Valence was coded separately for e-cigarettes and other tobacco products. Tweets/videos that on the whole were mostly supportive of using the product or the industry making the product were coded “Pro”, while tweets/videos that were mostly against the use of the product or the industry
making the product were coded “Anti”. For Twitter, precision varied .69 to .92, and recall varied .53 to .94. For YouTube tobacco valence, precision was .60 for anti and .96 for pro; recall was .93 for both. Since the vast majority of manually labeled videos (97%) were pro-ecig, an ecig valence classifier was not built.

Four tobacco control themes were chosen a priori based on prior tobacco content analyses. We expected these themes would likely have an impact on the general public’s beliefs about tobacco products. (1) “Health”: effects of product use on the user’s physical health (specifically excluding health effects of use for non-users (e.g., secondhand smoke)), (2) “Policy”: mandatory policy/law/regulation by a government, company, or institution, (3) “Addiction”: explicit references to products being addictive or users being addicted, and (4) “Youth”: use, access, or purchase of products by anyone up to 21 years old. For Twitter, all four themes had precision and recall >.73. For YouTube, precision varied .59 to .82 and recall varied .33 to .55.

Like valence, normative information was also coded separately for e-cigarettes and other tobacco products. Normative information was divided into two types: “Population” and “Individual”. Population-level summary information (“population norms”) conveys behavior prevalence in an explicit and straightforward way by providing numbers, percentages or trends from census data, survey results, reports, or educational campaigns. None of these sources had enough population norm information to develop reliable classifiers for Twitter and YouTube, so these datasets were not coded for that. Individual-level behavior cues (“individual norms”), on the other hand, facilitate formation of descriptive norm perceptions more subtly by providing knowledge about and exposure to instances of others’ performing (or not performing) the behavior. Individual norms were coded for whether or not they described ecig or tobacco “use” (as opposed to a reduction in use, or non-use). There were not enough examples in our hand-coded samples to build classifiers for non-use, so the database was not coded for that either. For Twitter, precision was .52 for both ecig and tobacco, and recall was .74 for ecig and .71 for tobacco. For YouTube, precision was .82 for ecig and .88 for tobacco, and recall was .94 for ecig and .97 for tobacco.

**Current Set of Examples**

In all there are 10 different codes. In order to have some illustrative examples of the items that were coded, we pulled 15 examples for each valence, theme, and norm from Twitter and YouTube. These sets were pulled separately for tobacco (no ecig)/ecig items if they existed for that code. Also, in each set of 15, we chose 10 examples from Twitter and 5 examples from YouTube. In all there are 200 examples included here. Individual items were randomly selected from all items matching those criteria.

Hand-coding of the examples does not perfectly match the machine coding. This is expected because the test set precision and recall were not perfect (see above). Also, some sets don’t match for themes because the validity was tested across products (e.g., ecig addiction has several mistakes because the theme refers to a different product than the one indicated). Rather than the definitions of each code, the examples provided here are meant to give the reader a sense of the actual tweets/videos included for each code. We have included the team’s hand-coding of whether or not these examples match the code (Yes or No) before each example. Effects analyses aggregate the data to at least the day-level ensuring that the amount of valence, theme, and norm coverage is estimated from many individual tweets/videos, thus we are less concerned about individual tweets/videos being misclassified.
Table of Contents

How to use these examples ........................................................................................................... 5
Tobacco (no ecig) Pro ..................................................................................................................... 6
  Tobacco (no ecig) tweets/videos ............................................................................................... 6
Tobacco (no ecig) Anti ................................................................................................................... 6
  Tobacco (no ecig) tweets/videos ............................................................................................... 6
Ecig Pro ........................................................................................................................................ 7
  E-cigarette tweets/videos ........................................................................................................ 7
Ecig Anti ....................................................................................................................................... 7
  E-cigarette tweets .................................................................................................................... 7
Health ......................................................................................................................................... 7
  Tobacco (no ecig) tweets/videos ............................................................................................... 7
  E-cigarette tweets/videos ........................................................................................................ 8
Policy .......................................................................................................................................... 8
  Tobacco (no ecig) tweets/videos ............................................................................................... 8
  E-cigarette tweets/videos ........................................................................................................ 9
Addiction ..................................................................................................................................... 9
  Tobacco (no ecig) tweets/videos ............................................................................................... 9
  E-cigarette tweets/videos ........................................................................................................ 10
Youth ......................................................................................................................................... 10
  Tobacco (no ecig) tweets/videos ............................................................................................... 10
  E-cigarette tweets/videos ........................................................................................................ 11
Tobacco (no ecig) Individual Norms ......................................................................................... 11
  Tobacco (no ecig) tweets/videos ............................................................................................... 11
Ecig Individual Norms ............................................................................................................... 12
  E-cigarette tweets/videos ........................................................................................................ 12
How to use these examples

This document uses Word’s heading styles to make the examples easier to navigate. Each heading style can be opened or closed to reveal content lower down. All examples are closed by default when opening the document. No matter which examples you open, they will be closed the next time you open Word. Instructions for expanding all headings are available here: https://support.office.com/en-us/article/collapse-or-expand-parts-of-a-document-701786e0-95e2-40bf-bfe5-f0233cd3520c

To open an example:
  • Hover your cursor over the blue labels and an open triangle will appear to the left of each label.
  • Click the open triangle to open the tweet/video.

To close an example:
  • Hover your cursor over the blue labels again and a closed triangle will appear to the left of each label. Click the closed triangle to close the example. You can also close entire sections using the same method.
Tobacco (no ecig) Pro
Note: For analysts of the dataset, the code for Tobacco (no ecig) Pro within the database for Twitter is “twtpro”. The YouTube code is “pro_tob”.

Tobacco (no ecig) tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_tpro”.

Tobacco (no ecig) Pro – Twitter 1
Tobacco (no ecig) Pro – Twitter 2
Tobacco (no ecig) Pro – Twitter 3
Tobacco (no ecig) Pro – Twitter 4
Tobacco (no ecig) Pro – Twitter 5
Tobacco (no ecig) Pro – Twitter 6
Tobacco (no ecig) Pro – Twitter 7
Tobacco (no ecig) Pro – Twitter 8
Tobacco (no ecig) Pro – Twitter 9
Tobacco (no ecig) Pro – Twitter 10
Tobacco (no ecig) Pro – YouTube 1
Tobacco (no ecig) Pro – YouTube 2
Tobacco (no ecig) Pro – YouTube 3
Tobacco (no ecig) Pro – YouTube 4
Tobacco (no ecig) Pro – YouTube 5

Tobacco (no ecig) Anti
Note: For analysts of the dataset, the code for Tobacco (no ecig) Anti within the database for Twitter is “twtanti”. The YouTube code is “anti_tob”.

Tobacco (no ecig) tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_tanti”.

Tobacco (no ecig) Anti – Twitter 1
Tobacco (no ecig) Anti – Twitter 2
Tobacco (no ecig) Anti – Twitter 3
Tobacco (no ecig) Anti – Twitter 4
Tobacco (no ecig) Anti – Twitter 5
Tobacco (no ecig) Anti – Twitter 6
Tobacco (no ecig) Anti – Twitter 7
Tobacco (no ecig) Anti – Twitter 8
Tobacco (no ecig) Anti – Twitter 9
Tobacco (no ecig) Anti – Twitter 10
Tobacco (no ecig) Anti – YouTube 1
Tobacco (no ecig) Anti – YouTube 2
Tobacco (no ecig) Anti – YouTube 3
Tobacco (no ecig) Anti – YouTube 4
Tobacco (no ecig) Anti – YouTube 5

Ecig Pro
Note: For analysts of the dataset, the code for Ecig Pro within the database for Twitter is “twepro”.

E-cigarette tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “ecig_tweet_epro”.

Ecig Pro – Twitter 1
Ecig Pro – Twitter 2
Ecig Pro – Twitter 3
Ecig Pro – Twitter 4
Ecig Pro – Twitter 5
Ecig Pro – Twitter 6
Ecig Pro – Twitter 7
Ecig Pro – Twitter 8
Ecig Pro – Twitter 9
Ecig Pro – Twitter 10
Ecig Pro – YouTube

Ecig Anti
Note: For analysts of the dataset, the code for Ecig Anti within the database for Twitter is “tweanti”.

E-cigarette tweets
Note: For analysts of the dataset, the variable name within the dataset is “ecig_tweet_eanti”.

Ecig Anti – Twitter 1
Ecig Anti – Twitter 2
Ecig Anti – Twitter 3
Ecig Anti – Twitter 4
Ecig Anti – Twitter 5
Ecig Anti – Twitter 6
Ecig Anti – Twitter 7
Ecig Anti – Twitter 8
Ecig Anti – Twitter 9
Ecig Anti – Twitter 10

Health
Note: For analysts of the dataset, the code for Health within the database for Twitter is “twhea”. The
YouTube code is “health_pred”.

Tobacco (no ecig) tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_hea”.
Health – Tobacco (no ecig) – Twitter 1
Health – Tobacco (no ecig) – Twitter 2
Health – Tobacco (no ecig) – Twitter 3
Health – Tobacco (no ecig) – Twitter 4
Health – Tobacco (no ecig) – Twitter 5
Health – Tobacco (no ecig) – Twitter 6
Health – Tobacco (no ecig) – Twitter 7
Health – Tobacco (no ecig) – Twitter 8
Health – Tobacco (no ecig) – Twitter 9
Health – Tobacco (no ecig) – Twitter 10
Health – Tobacco (no ecig) – YouTube 1
Health – Tobacco (no ecig) – YouTube 2
Health – Tobacco (no ecig) – YouTube 3
Health – Tobacco (no ecig) – YouTube 4
Health – Tobacco (no ecig) – YouTube 5

E-cigarette tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “ecig_SOURCE_hea”.
Health – Ecig – Twitter 1
Health – Ecig – Twitter 2
Health – Ecig – Twitter 3
Health – Ecig – Twitter 4
Health – Ecig – Twitter 5
Health – Ecig – Twitter 6
Health – Ecig – Twitter 7
Health – Ecig – Twitter 8
Health – Ecig – Twitter 9
Health – Ecig – Twitter 10
Health – Ecig – YouTube 1
Health – Ecig – YouTube 2
Health – Ecig – YouTube 3
Health – Ecig – YouTube 4
Health – Ecig – YouTube 5

Policy
Note: For analysts of the dataset, the code for Policy within the database for Twitter is “twpol”. The YouTube code is “policy_pred”.
Tobacco (no ecig) tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_pol”.

Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_pol”.
Policy – Tobacco (no ecig) – Twitter 1
Policy – Tobacco (no ecig) – Twitter 2
Policy – Tobacco (no ecig) – Twitter 3
Policy – Tobacco (no ecig) – Twitter 4
Policy – Tobacco (no ecig) – Twitter 5
Policy – Tobacco (no ecig) – Twitter 6
Policy – Tobacco (no ecig) – Twitter 7
Policy – Tobacco (no ecig) – Twitter 8
Policy – Tobacco (no ecig) – Twitter 9
Policy – Tobacco (no ecig) – Twitter 10
Policy – Tobacco (no ecig) – YouTube 1
Policy – Tobacco (no ecig) – YouTube 2
Policy – Tobacco (no ecig) – YouTube 3
Policy – Tobacco (no ecig) – YouTube 4
Policy – Tobacco (no ecig) – YouTube 5

E-cigarette tweets/videos

Note: For analysts of the dataset, the variable name within the dataset is “ecig_SOURCE_pol”.

Policy – Ecig – Twitter 1
Policy – Ecig – Twitter 2
Policy – Ecig – Twitter 3
Policy – Ecig – Twitter 4
Policy – Ecig – Twitter 5
Policy – Ecig – Twitter 6
Policy – Ecig – Twitter 7
Policy – Ecig – Twitter 8
Policy – Ecig – Twitter 9
Policy – Ecig – Twitter 10
Policy – Ecig – YouTube 1
Policy – Ecig – YouTube 2
Policy – Ecig – YouTube 3
Policy – Ecig – YouTube 4
Policy – Ecig – YouTube 5

Addiction

Note: For analysts of the dataset, the code for Addiction within the database for Twitter is “twadd”. The YouTube code is “addiction_pred”.

Tobacco (no ecig) tweets/videos

Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_add”.

Addiction – Tobacco (no ecig) – Twitter 1
Addiction – Tobacco (no ecig) – Twitter 2
Addiction – Tobacco (no ecig) – Twitter 3
Addiction – Tobacco (no ecig) – Twitter 4
Addiction – Tobacco (no ecig) – Twitter 5
Addiction – Tobacco (no ecig) – Twitter 6
Addiction – Tobacco (no ecig) – Twitter 7
Addiction – Tobacco (no ecig) – Twitter 8
Addiction – Tobacco (no ecig) – Twitter 9
Addiction – Tobacco (no ecig) – Twitter 10
Addiction – Tobacco (no ecig) – YouTube 1
Addiction – Tobacco (no ecig) – YouTube 2
Addiction – Tobacco (no ecig) – YouTube 3
Addiction – Tobacco (no ecig) – YouTube 4
Addiction – Tobacco (no ecig) – YouTube 5
E-cigarette tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “ecig_SOURCE_add”.

Addiction – Ecig – Twitter 1
Addiction – Ecig – Twitter 2
Addiction – Ecig – Twitter 3
Addiction – Ecig – Twitter 4
Addiction – Ecig – Twitter 5
Addiction – Ecig – Twitter 6
Addiction – Ecig – Twitter 7
Addiction – Ecig – Twitter 8
Addiction – Ecig – Twitter 9
Addiction – Ecig – Twitter 10
Addiction – Ecig – YouTube 1
Addiction – Ecig – YouTube 2
Addiction – Ecig – YouTube 3
Addiction – Ecig – YouTube 4
Addiction – Ecig – YouTube 5

Youth
Note: For analysts of the dataset, the code for Youth within the database for Twitter is “twyou”. The YouTube code is “youth_pred”.

Tobacco (no ecig) tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_you”.
Youth – Tobacco (no ecig) – Twitter 1
Youth – Tobacco (no ecig) – Twitter 2
Youth – Tobacco (no ecig) – Twitter 3
Youth – Tobacco (no ecig) – Twitter 4
Youth – Tobacco (no ecig) – Twitter 5
Youth – Tobacco (no ecig) – Twitter 6
Youth – Tobacco (no ecig) – Twitter 7
Youth – Tobacco (no ecig) – Twitter 8
Youth – Tobacco (no ecig) – Twitter 9
Youth – Tobacco (no ecig) – Twitter 10
Youth – Tobacco (no ecig) – YouTube 1
Youth – Tobacco (no ecig) – YouTube 2
Youth – Tobacco (no ecig) – YouTube 3
Youth – Tobacco (no ecig) – YouTube 4
Youth – Tobacco (no ecig) – YouTube 5

E-cigarette tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “ecig_SOURCE_you”.
Youth – Ecig – Twitter 1
Youth – Ecig – Twitter 2
Youth – Ecig – Twitter 3
Youth – Ecig – Twitter 4
Youth – Ecig – Twitter 5
Youth – Ecig – Twitter 6
Youth – Ecig – Twitter 7
Youth – Ecig – Twitter 8
Youth – Ecig – Twitter 9
Youth – Ecig – Twitter 10
Youth – Ecig – YouTube 1
Youth – Ecig – YouTube 2
Youth – Ecig – YouTube 3
Youth – Ecig – YouTube 4
Youth – Ecig – YouTube 5

Tobacco (no ecig) Individual Norms
Note: For analysts of the dataset, the code for Tobacco (no ecig) Individual Norms within the database for Twitter is “twtindnorm”. The YouTube code is “tob_pu”.
Tobacco (no ecig) tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “tobnoecig_SOURCE_tindnorm”.
Tobacco (no ecig) Individual Norms – Twitter 1
Tobacco (no ecig) Individual Norms – Twitter 2
Ecig Individual Norms
Note: For analysts of the dataset, the code for Ecig Individual Norms within the database for Twitter is “tweindnorm”. The YouTube code is “ecig_pu”.

E-cigarette tweets/videos
Note: For analysts of the dataset, the variable name within the dataset is “ecig_SOURCE_eindnorm”.

Ecig Individual Norms – Twitter 1
Ecig Individual Norms – Twitter 2
Ecig Individual Norms – Twitter 3
Ecig Individual Norms – Twitter 4
Ecig Individual Norms – Twitter 5
Ecig Individual Norms – Twitter 6
Ecig Individual Norms – Twitter 7
Ecig Individual Norms – Twitter 8
Ecig Individual Norms – Twitter 9
Ecig Individual Norms – Twitter 10
Ecig Individual Norms – YouTube 1
Ecig Individual Norms – YouTube 2
Ecig Individual Norms – YouTube 3
Ecig Individual Norms – YouTube 4
Ecig Individual Norms – YouTube 5