Towards Sentiment Analysis of Tobacco Products' Usage in Social Media

Venkata Himakar Yanamandra [*], **Kartikey Pant** [*] and **Radhika Mamidi**
International Institute of Information Technology, Hyderabad, India
{himakar.y, kartikey.pant}@research.iiit.ac.in,
radhika.mamidi@iiit.ac.in.

Abstract

Contemporary tobacco-related studies are mostly concerned with a single social media platform while missing out on a broader audience. Moreover, they are heavily reliant on labeled datasets, which are expensive to make. In this work, we explore sentiment and product identification on tobacco-related text from two social media platforms. We release *SentiSmoke-Twitter* and *SentiSmoke-Reddit* datasets, along with a comprehensive annotation schema for identifying tobacco products' sentiment. We then perform benchmarking text classification experiments using state-of-the-art models, including *BERT*, *RoBERTa*, and *DistilBERT*. Our experiments show F1 scores as high as 0.72 for sentiment identification in the Twitter dataset, 0.46 for sentiment identification, and 0.57 for product identification using semi-supervised learning for Reddit.

1 Introduction

Smoking tobacco causes more than 8 million deaths each year[1]. While cigarette smoking has fallen globally, e-cigarettes remain the commonly used tobacco product among the US youth[2].

So far, 2,807 cases and 68 deaths have been recorded due to e-cigarette use-associated lung injury in the US[2] alone. These developments prompted the US government to enforce a countrywide ban on addictive flavors and sales to minors. As smokers are more likely to develop severe COVID-19 symptoms, WHO has introduced the world's first digital health worker *Florence* based on AI to fill the gap between overburdened medical system users who are trying to quit[3]. Health-related discourses on social media have risen substantially over the years (Tamersoy et al., 2015) since social media provides an opportunity to informally express opinions freely with like-minded individuals across geographical and societal barriers.

Twitter boasts of 330 million active monthly users who send out half a billion tweets daily[4] while Reddit has 430 million monthly users and has $100k+$ communities [5]. Reddit provides expendable and pseudo-anonymous accounts that are well suited for controversial discussions, including the perception of electronic cigarettes and marijuana, which might be inappropriate to discuss on non-anonymous forums (Park and Conway, 2018).

In Salathé and Khandelwal (2011), we see that information flows more often between users who share the same sentiments. Studying trends of first-person accounts of experience and sentiment towards different tobacco products in these forums becomes imperative in the ongoing bio-surveillance and regulatory efforts (Kim et al., 2015; Pant et al., 2019). Real-time monitoring of public sentiment and informativeness gives us an opportunity for bottom-up discovery of emergent patterns, especially in vulnerable and ethnically diverse populations that may not be readily detectable by traditional methods (Myslín et al., 2013; Lienemann et al., 2017). Although Yanamandra et al. (2020) identified tobacco products on Twitter, they leave a gap in identifying the products' sentiment.

This work explores tobacco products' identification and their sentiments in texts extracted from two commonly-used social media platforms. We release two datasets for multiclass classification annotated with tobacco products and their sentiment: *SentiSmoke-Twitter* and *SentiSmoke-Reddit*. [6] We

[*] The first two authors have contributed equally to the work.

[1] http://bit.ly/WHO_tob

[2] https://www.cdc.gov/tobacco/

[3] https://bit.ly/who_AI

[4] http://bit.ly/tweet_stats

[5] https://www.redditinc.com/

[6] https://github.com/himakaryv/SentiSmoke-Datasets

Proceedings of Recent Advances in Natural Language Processing, pages 1545–1552
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_173

Positive	*Twitter*	"Cigarette a day keeps the depression away"
	Reddit	"I just wanted to say that vaping has saved my life"
Negative	*Twitter*	" When girls post a photo but have a cigarette in their hand #putoff #notcool "
	Reddit	"Day 16 of quitting, just got rid of my vape"
Neutral	*Twitter*	"You are my nicotine, heroine, novocaine. Fvck. You're mah drug. Im addicted to you."
	Reddit	"Tobacco stocks rise as FDA delays plan to cut nicotine levels in cigarettes"

Table 1: Examples from the Dataset for each class and source.

further benchmark state-of-the-art text classification models for the supervised and semi-supervised learning tasks of identifying product and sentiment in Twitter and Reddit. To the best of our knowledge, this is the first cross-platform semi-supervised attempt in tobacco research.

2 Related Work

There has been considerable work in the field of health and social media. While Salathé and Khandelwal (2011) explored Spatio-temporal sentiment towards the new influenza vaccine in Twitter, Cole et al. (2016) studied the quality characteristics of information found in online health forums like Reddit, Mumsnet, and Patient covering HIV, diabetes, and chickenpox. Moreover, Park and Conway (2018) tracked public interest in Ebola, e-cigarettes, influenza, and marijuana on Reddit.

Recent academic works have analyzed tobacco-related trends in multiple social media platforms, including Twitter, Reddit, Instagram, and YouTube (Allem et al., 2019; Carroll et al., 2012; Yanamandra et al., 2020; Zhang et al., 2018). While Allem et al. (2018) explored a thematic analysis of hookah Twitter posts, Allem et al. (2017a) explored e-cigarette trends with an emphasis on social bot behaviour. In Kim et al. (2015), an infoveillance study was performed on e-cigarette tweets on the themes of marketing and usage locations. Moreover, Chen et al. (2015) compared consumer experiences across online discussion forums, including Reddit. They focused on hookah, e-cigarette, and cigarettes using topic modelling and visualization. Sharma et al. (2016) performed a qualitative thematic analysis to determine the limitations and motivations for e-cigarette users with mental illness on Reddit.

The analysis of tobacco sentiment analysis is a relatively less-explored problem. In Allem et al. (2017b), the authors conducted the sentiment analysis of hookah-related tweets using SVM. Moreover, Myslín et al. (2013) performed content and sentiment analysis on tobacco-related tweets using Naive Bayes, KNN, and SVM.

3 Dataset

We use the *SmokPro* (Yanamandra et al., 2020) dataset, which consists of $2,116$ tobacco-related tweets classified into the following five distinct product classes: traditional tobacco product mention, modern tobacco product mention, general mention of smoking, narcotics & other drug mentions, and ambivalent mentions. The authors consider *cigarette, hookah, pipe, cigar, bidis, cigarillo, shisha,* and *baccy* as traditional tobacco products. Moreover, they consider *e-cigarette, e-juice, e-hookahs, e-liquid, mods, vape pens, vapes, tank systems,* and *electronic nicotine delivery systems (ENDS)* as modern tobacco products.

stopsmoking	*cigarettes*	*juul*	*vaping101*
vape	*tobacco*	*stonerprotips*	*ecigclassifieds*
quit vaping	*hookah*	*vaporents*	*DIY_eJuice*
vaping	*cigars*	*weed*	*vapeporn*
electronic_cigarette	*smokingcessation*	*leaves*	*nicotine*

Table 2: List of subreddits used to scrape data for SentiSmoke-Reddit.

Table 3 illustrates the schema used to determine each product's sentiment. Our guidelines are centered around the type of tobacco product based on the content. We have also considered street terms and colloquial slangs associated with tobacco product usage. We compile a list of 20 subreddits to scrape data that contained tobacco-related content. They include both information and cessation platforms and are listed in Table 2.

Using PRAW [7], we scrape a sample of Reddit posts made in the last year. We use Reddit's *Top* filter and only consider posts with more than 10 upvotes to reduce spam and fewer interacted-with posts. While Twitter's character limit is 280, Reddit's maximum character limits for the title and the body are 300 and $40,000$ characters, respectively. To avoid cross-platform discrepancies, we only consider the post's title for this task due to its

[7] https://github.com/praw-dev/praw

Positive	Negative	Neutral
General Usage	Past Usage Before Quitting	Sarcastic or Unclear Usage
Positive Experience of the Product	Negative Experience of the Product	
Cravings	Cravings experienced during current/past journey of Quitting	
Depicts the usage by a second-person or third-person	Shows antipathy towards the product and users	
Advertisements praising the product	Advertisements showcasing negative aspects or rehabilitation products	General Advertisements
Product reviews on various brands and flavours	Studies and statistics showcasing negative aspects and addictive nature	The tobacco product is not the focus of the text.
Campaigns supporting usage and condemning regulatory steps	Quitting and Cessation Campaigns and Movements	Usage of smoking-related words in other contexts. For example, "smoking gun", "smoking hot".
Complaining about minor inconveniences occurring due to tobacco-use	Indicates that the interlocutor is acting absurdly by comparing to being intoxicated. For instance, "What are you smoking?"	Other products and places named after smoking-related terms. For example, "cigarette pants", "tobacco docks".

Table 3: The data annotation schema used for Tobacco Product Sentiment Analysis.

similar character length compared with the tweets. Moreover, the effect of title in a Reddit post has been previously studied by Horne and Adali (2017). The scraped dataset consists of $10,023$ titles.

We divide the dataset into three splits: *training*, consisting of $9,023$ titles, *val*, consisting of 100 titles, *test*, consisting of 900 titles. Since we utilize semi-supervised learning for this task, we only annotate the *val*, and *test* splits to determine the quality of the learning process. We annotate product labels using the *SmokPro* 's product annotation schema. We then use the aforementioned sentiment schema (Table 3) for annotating sentiment labels. The annotation process was done by two human annotators having a fluent English background. To assess the annotation standard, we calculate the Inter-Annotator Agreement(IAA) using Cohen's Kappa coefficient (Fleiss and Cohen, 1973). We obtain Kappa Scores of 0.896 for sentiment annotation for the *SentiSmoke-Twitter* dataset and 0.840, and 0.927 for sentiment and product identification for the *SentiSmoke-Reddit* dataset, respectively. These indicate the high quality of all three annotation processes.

4 Methodology

This section briefly describes the methodology used for our benchmarking experiments and details the semi-supervised text classification for the Reddit-based dataset.

4.1 Models Used

4.1.1 FastText

FastText (Joulin et al., 2016) is an open-source, free, lightweight NLP library. It maintains a memory-efficient mapping of n-grams and shares information across classes through a hidden representation. FastText uses a hashtable for the word and character n-grams, with the hashtable size directly impacting the size of a model. This library's accuracy has been found on par with deep neural networks while requiring a fractional amount of training time.

4.1.2 BERT

BERT (Devlin et al., 2019) is a contextualized language representation model based on bidirectional transformers. It uses novel pre-training objectives like masked-language modelling and next-sentence prediction, which enhance the modelling of a relationship between two sentences. In masked-language modelling, the model randomly masks a random subset of the input tokens, and the objective is to predict the correct tokens based purely on context. On the other hand, the next-sentence prediction simultaneously pre-trains text-pair representations. These features help BERT outperform previous state-of-the-art techniques by a large margin. It uses word-piece tokenization and embeddings, which splits parts of words to get better word information and decreases overall vocabulary

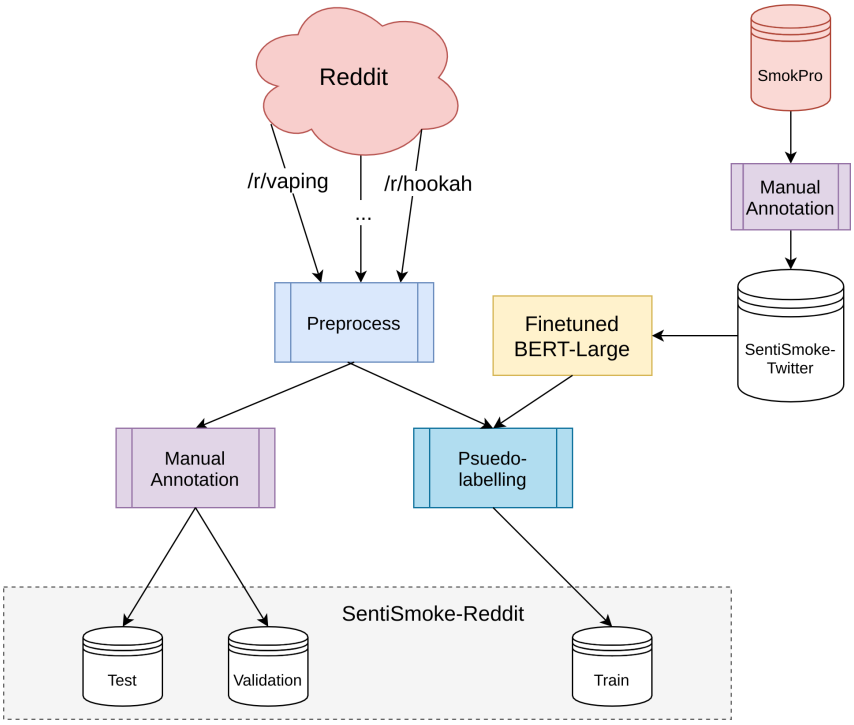

Figure 1: Methodology diagram for creating SentiSmoke-Twitter and SentiSmoke-Reddit.

size effectively. We benchmark *cased* $BERT_{Base}$ and *cased* $BERT_{Large}$ and fine-tune them for our classification experiments.

4.1.3 RoBERTa

RoBERTa (Liu et al., 2019) is a BERT-based model with improved training methods, larger training data size, and higher computational power. It is trained on ten times the training data as BERT. RoBERTa has been used in multiple online NLP studies published in the last few years in areas including disclosure modelling (Dadu et al., 2020).

In the improved training methodology, dynamic masking replaces BERT's next-sentence prediction. In dynamic masking, masked tokens change in between epochs. RoBERTa uses larger byte-pair encoding(BPE) vocabulary compared to BERT. These changes led RoBERTa to outperform BERT on GLUE benchmarks. We use the large variant of RoBERTa, $RoBERTa_{Large}$, and fine-tune it for our experiments.

4.1.4 DistilBERT

DistilBERT [8] is another BERT-based model utilizing knowledge distillation leading to a much smaller and faster model. It uses 40% less parameters than *Uncased* $BERT_{base}$, runs 60% faster and preserves 95% of *Uncased* $BERT_{base}$'s performance measured on GLUE benchmark. Distilled models have been previously used in downstream tasks with a good predictive performance including subjective bias detection (Pant et al., 2020). We fine-tune the pre-trained *cased base* variant of the model for our experiments.

4.2 Transfer Learning

We exploit pseudo-labelling to use the unannotated data scraped from Reddit to enhance the process of cross-platform transfer from Twitter to Reddit. For this task, we utilize the predictions from the best-performing model for each subtask. We then use the pseudo-labelled corpus and splits of the original Twitter-based datasets to predict labels for the Reddit evaluation split. This semi-supervised learning methodology is used to exploit the learnings from both Reddit and Twitter.

5 Experiments

In this section, we describe the text classification experiments performed using both supervised and semi-supervised learning. We also highlight the experimental settings along with the experimental results for each of the three experiments.

We conduct three experiments with the aforementioned datasets:

[8] https://huggingface.co/transformers/model_doc/distilbert.html

Methods	Sentiment Classification			
	Accuracy	**F1**	**Precision**	**Recall**
FastText	66.03%	0.655	0.661	0.660
Uncased $BERT_{Large}$	69.81%	0.695	0.695	0.698
Cased $BERT_{Large}$	68.39%	0.676	0.679	0.683
$RoBERTa_{Large}$	**72.64%**	**0.724**	**0.730**	**0.726**
DistilBERT	65.56%	0.650	0.661	0.655

Table 4: Experimental results for the *SentiSmoke-Twitter* test split.

Methods	Product Classification				Sentiment Classification			
	Accuracy	**F1**	**Precision**	**Recall**	**Accuracy**	**F1**	**Precision**	**Recall**
FastText	49.00%	0.502	0.622	0.490	43.70%	0.383	0.512	0.437
Uncased $BERT_{Large}$	48.13%	0.487	0.566	0.481	48.50%	0.445	0.549	0.485
Cased $BERT_{Large}$	55.48%	0.562	0.647	0.554	**49.00%**	0.448	0.550	**0.490**
$RoBERTa_{Large}$	**56.79%**	**0.573**	**0.667**	**0.567**	48.70%	**0.456**	**0.551**	0.487
DistilBERT	54.98%	0.557	0.642	0.549	47.40%	0.439	0.530	0.474

Table 5: Experimental results for the *SentiSmoke-Reddit* test split.

Sentiment Identification in Twitter is a supervised experiment that entails predicting the text's sentiment in the manually annotated *SentiSmoke-Twitter* dataset.

Product Identification in Reddit is a semi-supervised experiment that entails predicting the tobacco product of the text in the *SentiSmoke Reddit* dataset. We use the cased variant of $BERT_{Large}$, which had the highest predictive performance in the Twitter dataset (Yanamandra et al., 2020), to pseudo-label the dataset as described in Subsection 4.2.

Sentiment Identification in Reddit is another semi-supervised experiment that entails predicting the sentiment of the text in the *SentiSmoke-Reddit* dataset. We use $RoBERTa_{Large}$ to pseudo-label the dataset since it performed the best on Twitter.

We evaluate all the models on the following metrics: *F1*, *Precision*, *Recall*, and *Accuracy*. Moreover, for FastText, we use its automatic hyperparameter optimization functionality and validate it for 100 validation trials. For all BERT-based models and their distilled variants, we use a learning rate of $1 * 10^{-5}$ with a weight decay of 0.01, and an *adam epsilon* value of $1 * 10^{-8}$ while fine-tuning the models. We use a maximum sequence length of 100 and fine-tune the models for 2 epochs.

5.1 Results

We observe that $RoBERTa_{Large}$ outperforms all other models for all metrics for all the three experiments. From Table 4, we see that it obtains a high F1 score of 0.724 for the *Sentiment Identification in Twitter (Supervised)* task.

POS Tags	SentiSmoke-Reddit	SentiSmoke-Twitter
PUNCT	6.069	6.408
NOUN	2.850	3.135
VERB	1.885	2.258
PROPN	1.495	1.870
DET	1.472	1.304
ADP	1.246	1.168

Table 6: Average number of tokens per instance belonging to Top 5 most-occuring part-of-speech.

As illustrated in Table 5, experimental results show that $RoBERTa_{Large}$ again outperformed all other models for all metrics getting an F1 score of 0.573 for the *Product Identification in Reddit (Semi-Supervised)* task. For the *Sentiment Identification in Reddit* task, we see that $RoBERTa_{Large}$ outperformed all other models in F1 and Precision, scoring 0.456 and 0.551 on both metrics, respectively. On the other hand, Cased $BERT_{Large}$ obtains the highest Accuracy of 49.00% and Recall of 0.490. The performance of the models in the semi-supervised domain shows that the inductive transfer from Twitter to Reddit setting was effective for the tobacco-product-identification and sentiment-identification task.

For all three experiments, we infer that DistilBERT performs competitively with the large variant of its undistilled counterpart while taking the significantly lower time and computation power for the process of pre-training and fine-tuning.

6 Discussion

In the year 2016 alone, an estimated 10.5 million US youth were exposed to e-cigarette advertise-

Figure 2: Twitter Word Cloud.

Figure 3: Reddit Word Cloud.

ments through the internet [9]. CDC and FDA also discourages e-cigarette-related purchases through informal sources like friends or online marketplaces and forums [10]. Reddit was one of these online forums aforementioned where sales of tobacco products like e-cigarettes happened until the recent policy update prohibiting the sale of controlled substances like guns and drugs [11]. Therefore, we find it necessary to understand tobacco-related discussions and public sentiment to help shape better policies aimed at bio-surveillance and tobacco control measures.

Previous research (Benson et al., 2020; Pant et al., 2019) in sentiment analysis, and topical tobacco research is heavily reliant on manually annotated datasets. These pose several challenges - expensive to make, limited in scope, difficult to modify, and harder to scale. Additionally, previous research in this sphere is mostly concentrated only on a single social media platform. Both Reddit and Twitter have similar user demographics and a comparable number of monthly active users with a vibrant discourse on health-related issues. Our cross-platform supervised learning approach helps solve data scarcity and scalability while leveraging insights and context from one platform to another. We can see similarities of frequently used words like *smoking,vaping,weed,cigar* in word clouds generated for *SentiSmoke-Twitter* in Figure 2 and *SentiSmoke-Reddit* in Figure 3.

Moreover, extracting data from moderated topic subreddits directly instead of general keyword search used in previous studies (Park and Conway, 2017, 2018) helps us access targeted tobacco-related discourse while weeding out spam and unrelated information.

We also perform a part-of-speech-based analysis for comparing between the SentiSmoke-Twitter and SentiSmoke-Reddit, illustrated in Table 6. We have used *spaCy*[12] for this task. We note a high degree of similarity between the two datasets in terms of part-of-speech.

7 Conclusion and Future Works

This work explored sentiment and product identification on tobacco-related text from two social media platforms: Twitter and Reddit. We released two datasets for multiclass classification annotated across two axes: tobacco product and sentiment. We utilized semi-supervised learning on Reddit text using manually annotated text for Twitter. We then perform benchmarking experiments for sentiment and product identification in Reddit and Twitter using commonly used text classification models like FastText, BERT, RoBERTa, and DistilBERT. We obtain F1 scores as high as 0.72 for supervised sentiment identification in Twitter text, using manually annotated data. Our semi-supervised experiments on product and sentiment identification in Reddit text, using features learned from the Twitter text, obtain the F1 scores of 0.57 and 0.46, respectively.

Future work may involve using the predicted information in recommender systems, expanding the tasks for other social media platforms, and exploring the use of metadata-derived information and comments for the task. Our study can also be extended to real-time monitoring and bio-surveillance tools for social media, which takes continuous inflow of unseen data.

[9] http://bit.ly/ecig_marketing
[10] http://bit.ly/ecig_onlinesale
[11] https://bit.ly/reddit_policy
[12] https://spacy.io/api/annotation#pos-tagging

References

Jon-Patrick Allem, Likhit Dharmapuri, Adam Leventhal, Jennifer Unger, and Tess Cruz. 2018. Hookah-related posts to twitter from 2017 to 2018: Thematic analysis. *Journal of Medical Internet Research*, 20:e11669.

Jon-Patrick Allem, Emilio Ferrara, Sree Priyanka Uppu, Tess Boley Cruz, and Jennifer B Unger. 2017a. E-cigarette surveillance with social media data: Social bots, emerging topics, and trends. *JMIR Public Health Surveill*, 3(4).

Jon-Patrick Allem, Anuja Majmundar, Likhit Dharmapuri, Jennifer Unger, and Tess Cruz. 2019. Insights on electronic cigarette products from reviews on the reddit forum. *Tobacco Prevention Cessation*, 5.

Jon-Patrick Allem, Jagannathan Ramanujam, Kristina Lerman, Kar-Hai Chu, Tess Boley Cruz, and Jennifer B Unger. 2017b. Identifying sentiment of hookah-related posts on twitter. *JMIR Public Health Surveill*, 3(4):e74.

Ryzen Benson, Mengke Hu, Annie T Chen, Subhadeep Nag, Shu-Hong Zhu, and Mike Conway. 2020. Investigating the attitudes of adolescents and young adults towards juul: Computational study using twitter data. *JMIR Public Health Surveill*, 6(3):e19975.

Mary Carroll, Ariel Shensa, and Brian Primack. 2012. A comparison of cigarette- and hookah-related videos on youtube. *Tobacco control*, 22.

Annie Chen, Shu-Hong Zhu, and Mike Conway. 2015. What online communities can tell us about electronic cigarettes and hookah use: A study using text mining and visualization techniques. *Journal of Medical Internet Research*, 17:e220.

Jennifer Cole, Chris Watkins, and Dorothea Kleine. 2016. Health advice from internet discussion forums: How bad is dangerous? *Journal of Medical Internet Research*, 18:e4.

Tanvi Dadu, Kartikey Pant, and Radhika Mamidi. 2020. Bert-based ensembles for modeling disclosure and support in conversational social media text.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

B. Horne and Sibel Adali. 2017. The impact of crowds on news engagement: A reddit case study. *ArXiv*, abs/1703.10570.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. In *EACL*.

Annice Kim, Timothy Hopper, Sean Simpson, James Nonnemaker, Alicea (Allie) Lieberman, Heather Hansen, Jamie Guillory, and Lauren Porter. 2015. Using twitter data to gain insights into e-cigarette marketing and locations of use: An infoveillance study. *Journal of Medical Internet Research*.

Brianna Lienemann, Jennifer Unger, Tess Cruz, and Kar-Hai Chu. 2017. Methods for coding tobacco-related twitter data: A systematic review. *Journal of Medical Internet Research*, 19:e91.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15:e174.

Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. 2020. Towards detection of subjective bias using contextualized word embeddings. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 75–76, New York, NY, USA. Association for Computing Machinery.

Kartikey Pant, Venkata Himakar Yanamandra, Alok Debnath, and Radhika Mamidi. 2019. Smokeng: Towards fine-grained classification of tobacco-related social media text. In *W-NUT@EMNLP*.

Albert Park and Mike Conway. 2017. Towards tracking opium related discussions in social media. *Online Journal of Public Health Informatics*, 9.

Albert Park and Mike Conway. 2018. Tracking health related discussions on reddit for public health applications. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:1362–1371.

Marcel Salathé and Shashank Khandelwal. 2011. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS computational biology*, 7:e1002199.

Ratika Sharma, Britta Wigginton, Carla Meurk, Pauline Ford, and Coral Gartner. 2016. Motivations and limitations associated with vaping among people with mental illness: A qualitative analysis of reddit discussions. *International Journal of Environmental Research and Public Health*, 14:7.

Acar Tamersoy, Munmun Choudhury, and Duen Horng Chau. 2015. Characterizing smoking and drinking abstinence from social media. volume 2015, pages 139–148.

Venkata Himakar Yanamandra, Kartikey Pant, and R. Mamidi. 2020. Smokpro: Towards tobacco product identification in social media text. In *SIIRH@ECIR*.

Youshan Zhang, Jon-Patrick Allem, Jennifer Beth Unger, and Tess Boley Cruz. 2018. Automated identification of hookahs (waterpipes) on instagram: An application in feature extraction using convolutional neural network and support vector machine classification. *J Med Internet Res*, 20(11):e10513.