2020

# Text Mining Methods for Analyzing Online Health Information and Communication

Sifei Han
*University of Kentucky*, sehan2@g.uky.edu

Digital Object Identifier: https://doi.org/10.13023/etd.2020.057

## Recommended Citation

Han, Sifei, "Text Mining Methods for Analyzing Online Health Information and Communication" (2020). *Theses and Dissertations--Computer Science*. 93.
https://uknowledge.uky.edu/cs_etds/93

Sifei Han, Student

Dr. Ramakanth Kavuluru, Major Professor

Dr. Mirosław Truszczyński, Director of Graduate Studies

# Text Mining Methods for Analyzing Online Health Information and Communication

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Engineering at the
University of Kentucky

By
Sifei Han
Lexington, Kentucky

Director: Dr. Ramakanth Kavuluru, Associate Professor of Biomedical Informatics
Lexington, Kentucky 2020

# ABSTRACT OF DISSERTATION

Text Mining Methods for Analyzing Online Health Information and Communication

The Internet provides an alternative way to share health information. Specifically, social network systems such as Twitter, Facebook, Reddit, and disease specific online support forums are increasingly being used to share information on health related topics. This could be in the form of personal health information disclosure to seek suggestions or answering other patients' questions based on their history. This social media uptake gives a new angle to improve the current health communication landscape with consumer generated content from social platforms. With these online modes of communication, health providers can offer more immediate support to the people seeking advice. Non-profit organizations and federal agencies can also diffuse preventative information in such networks for better outcomes. Researchers in health communication can mine user generated content on social networks to understand themes and derive insights into patient experiences that may be impractical to glean through traditional surveys. The main difficulty in mining social health data is in separating the signal from the noise. Social data is characterized by informal nature of content, typos, emoticons, tonal variations (e.g. sarcasm), and ambiguities arising from polysemous words, all of which make it difficult in building automated systems for deriving insights from such sources.

In this dissertation, we present four efforts to mine health related insights from user generated social data. In the first effort, we build a model to identify marketing tweets on electronic cigarettes (e-cigs) and assess different topics in marketing and non-marketing messages on e-cigs on Twitter. In our next effort, we build ensemble models to classify messages on a mental health forum for triaging posts whose authors need immediate attention from trained moderators to prevent self-harm. The third effort deals with models from our participation in a shared task on identifying tweets that discuss adverse drug reactions and those that mention medication intake. In the final task, we build a classifier that identifies whether a particular tweet about the popular Juul e-cig indicates the tweeter actually using the product. Our methods range from linear classifiers (e.g., logistic regression), classical nonlinear models (e.g., nearest neighbors), recent deep neural networks (e.g., convolutional neural networks), and ensembles of all these models in using different supervised training regimens (e.g., co-training). The focus is more on task specific system building than on building specific individual models. Overall, we demonstrate that it is possible to glean insights from social data on health related topics through natural language processing and machine learning with use-cases from substance use and mental health.

KEYWORDS: Natural language processing, machine learning, deep learning, neural networks, text classification, social data

Author's signature: Sifei Han

Date: February 25, 2020

Text Mining Methods for Analyzing Online Health Information and Communication

By
Sifei Han

Director of Dissertation: Ramakanth Kavuluru

Director of Graduate Studies: Mirosław Truszczyński

Date: February 25, 2020

Dedicated to my parents, Ying Han and Yalian Tang, my aunts, Lingyun Tang and Jennifer Tang, and my grandparents, Kaining Tang and Yinghua Peng.

# ACKNOWLEDGMENTS

First, I would like to thank my advisor, Dr. Ramakanth Kavuluru. I could not have finished my Ph.D. without his guidance and encouragement over the last several years. After I finished the undergraduate program, I didn't really know how to proceed further especially in a research direction. Then, I started my graduate studies and was fortunate to have Dr. Kavuluru to supervise my activities. As a newbie in the academic research setting at the time, it was critical to have his inputs to nudge me in the right direction and to better formulate research questions. As an international student and nonnative speaker of English, my writing unavoidably had some grammar and style issues. Dr. Kavuluru spent a significant amount of time to proofread and correct my mistakes. Overall his continuous support over the past few years was central to the process of successfully finishing and defending this dissertation.

I would also like to thank my committee members: Drs. Zongming Fei, Nathan Jacobs, and Giuseppe Labianca. Their advice and comments in the preparation of my dissertation have been invaluable. Likewise, I would also like to thank the Department of Computer Science that supported me financially during my Ph.D. study.

I also want to acknowledge my lab mates Anthony Rios, Zhiguo Yu, and Tung Tran, for the discussions on methodological topics. The advice I have received from them was vital to my success during my time in graduate school.

Last but not least, I would like to express my sincere gratitude to my family for their endless love and support to me throughout my time as a student.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

## Chapter 1 Introduction

In the last decade, social media platforms have shown rapid growth worldwide bringing a lot of new changes in people's daily lives. Online information sharing and consumption have become popular through social networks such as Facebook, Twitter, and Instagram and forums such as Reddit, Quora, and Stack Overflow. According to a 2015 survey by Pew Research Internet Project (Pew Research Internet Project, 2015), 76% online US adults use a social networking site. 71% of online adults use Facebook, while Twitter and Instagram are used by around 18% of them (Pew Research Internet Project, 2013c). One in four online US teenagers and 39% of African American online teenagers use Twitter (Pew Research Internet Project, 2013b). Teenagers favor Twitter and Instagram over Facebook (Piper Jaffray Market Research Project, 2014).

Twitter was introduced in 2006 and soon became one of the top ten websites (Alexa, 2016) on the Web with over 100 million daily active users who generate over 500 million tweets per day (Twitter, Inc, 2013). Twitter users (tweeters) can publicly express their opinion through short messages, called *tweets*, restricted to 140 characters. Tweeters can also engage in real-time conversations/discussion with other tweeters by *replying* or using a *hashtag* to join the conversations. Twitter is an asymmetric network, a tweeter can only receive the feed from the tweeters who he/she *follow*, and the tweeter won't receive their followers feeds until he/she *follow* back.

Reddit is an online forum that started in 2005 focusing on topics in entertainment, news, health, music, sports, gaming, and food. Reddit users (or Redditors) can submit Web links or text posts and receive *comments, upvotes*, and *down-votes.*To rank posts, Reddit uses a method based on the Newton cooling method and up/down votes. Unlike Twitter, Reddit allows longer, more detailed posts which can provide richer information in a better-structured format. In this proposal, we will focus on some subreddits (same as sub-forums/topical related discussion groups). While Twitter uses hashtags to track themes, Reddit uses subreddits to manage the themes.

As the content of social networks gets more in depth, it has changed the way researchers conduct their research and patients seek health information. Everyday social network users create millions of posts. Compared with telephone/volunteer-based survey methods to collect data, social network based data collection is an impossible mission for the human mind to analyze the posts manually. Identifying the different aspects of the themes that the posts describe, and evaluating the opinion of each aspect of the posts is computationally challenging. Researchers are developing automated methods to analyze these large, unstructured datasets. In this context machine learning, natural language processing (NLP), and statistical analyses present the possibility of utilizing this massive data for deriving insights from social streams.

Traditionally patients get information from health providers, while healthcare providers mainly focus on the clinical impact of the disease. Therefore, a patient's emotional side and social determinants of health might not get enough attention. With the social networks becoming more and more popular, they are not only used as social chatting platforms but also to share their health-related information on the general purpose social media (Twitter, Reddit, Facebook) and health orientated social networks (Lungevity, Cancer Survivor Network, EcigForum). This brings a new source for patients to seek information about their diseases and get social support.

In the health-orientated social networks, people sharing their health information give reviews about health providers and encourage each other in their fights with diseases. This is a complementary to the traditional face-to-face offline patients' networks. In the offline based patient's network, there exists several limitations: 1) many patients are physically weak (cannot walk, drive to the gathering location), 2) some patients have full-time jobs and other things they need to take care of, hindering them from meeting regularly, and 3) patients living in rural areas with rare diseases have difficulty finding groups. With the power of the Internet, people around the world can be easily connected and can instantly seek the information they need. From a researcher's perspective, there is growing evidence that there is enough signal in the social data to identify epidemic outbreaks in near real time, reach out to citizens for disaster relief, mine adverse drug reactions, and drug side effects. Also, mining disease specific forums can also enable providers, governmental organizations, and trained citizens in better communicating with healthcare consumers as a complementary means of delivering care and support.

## 1.1 Health Information

Information is facts provided about something or someone; therefore, health information is health related facts about something or someone. It contains patients' clinical context such as health history, diagnoses, allergies, current treatments, drug side effects, and health-related lifestyle factors (e.g., exercise, smoking, drinking). Usually, personal health records such as  (electronic medical records) are kept securely for view only by patients and health providers. As online health forums and social networks get more attention, it provides a new way for patients to seek information. By sharing their personal story, and viewing others' posts, patients can have a deeper understanding of what they may face in the future; how to find a good health provider; and where to find support groups.

As social media is not as regulated and is available for everyone to post their thoughts, we typically don't know who the posters are in real life or to what extent are they knowledgeable. The following concerns need to be considered:

- The trustworthiness of these posts
- Privacy issues
- Misleading information
- Incomplete data

From these health-related data, several types of health-related information can be extracted including

- Public beliefs, perceptions, and attitudes toward products, regulations (e.g., e-cigarettes, vaccination)
- Latest trends in substance abuse and addiction
- Side effects of newly introduced drugs in the market
- Factors associated with mental health concerns including suicide, depression, and anxiety.

## 1.2 Health Communication

Health communication traditionally is between doctors and patients and involves the doctor trying to change the patient's attitudes, external structures, and/or modify or eliminate certain behaviors (*What is Health Communications?* 2004). In this proposal, our definition is broader; anyone's discussion or post about health-related information counts as health communication which is not limited to patients and doctors. In the Online Health Communities (OHCs), people talk about their situations, and the repliers can be patients who are facing the same issue and share their suggestions or health providers giving care related suggestions.

In OHCs, various studies can be performed:

- Social support interventions
- Assistance for community facilitators
- Measurement of impact of participation

In traditional health studies, researchers from sociology and public health were using a small amount of data and manually analyzing it. In this dissertation, we particularly focus on facilitating interventions in mental health forums.

## 1.3 Thesis Statement

Analyzing online health information and communication is an important area in public health research. It is not realistic for humans to sift through millions of posts in social streams to derive health-related insights. Here we demonstrate that using NLP and machine learning methods with a small amount of human annotated data, we can extract health-related information and assist in moderating health communication in a timely manner.

## 1.4 Organization

The chapters in this dissertation are organized as follows:

**Chapter 2** introduces relevant related work shared among all the methods in this dissertation. This chapter address the notation used to describe the neural network methods throughout this manuscript. We then introduce the evaluation measures and present a general introduction to methods used in the rest of the dissertation.

**Chapter 3** represents a thematic analyses of electronic cigarette messages on Twitter. We created a binary classifier with accuracy $\approx 90\%$ to separate marketing tweets from regular tweets. Then we applied topic modeling to understand the themes of messages that belong to the two different groups. While marketing messages focused on sales, the non-marketing tweets contained messages across various subjects including e-cig side effects, burn injuries from exploding cartridges, and regulation stances.

**Chapter 4** deals with a small dataset situation with the use-case of triaging messages that need moderator attention in an online mental health support forum. We proposed an incremental undersampling technique to address the training data scarcity and employed ensemble models with feature selection to improve significantly over the prior best scores on the public dataset used.

**Chapter 5** describes a novel attention-based convolutional neural network architecture that helps the classifier focus on more important *n*-grams. It presents models for two subtasks from the 2017 Social Media Mining for Health(SMM4h) shared task on identifying adverse drug reactions and medication intake messages on Twitter. Our models are top performers for both tasks in the full dataset.

**Chapter 6** discusses a classifier that identifies if a tweet about the popular e-cig Juul indicates whether the author of the tweet uses it. The high level idea is to identify who uses Juul based on hints they provide on social media messages. This model can be used for downstream demographic studies, say for example to estimate (lower bound) underage consumption of Juul based on social media samples. While Chapter 4 uses traditional machine learning and Chapter 5 uses deep learning alone, combining traditional machine learning and deep neural networks outperform each individual model for this task.

**Chapter 7** concludes the dissertation by summarizing the important contributions and results. Our study shows that based on the goal of the task and the volume of data, we need to pick the model wisely. Deep models do not always outperform the traditional approaches.

## 1.5 Related Publications

This dissertation contains material previously published in the following papers:

- **Sifei Han**, Tung Tran, Anthony Rios, Ramakanth Kavuluru. "Team UKNLP: Detecting ADRs, Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter." In Proceedings of the 2nd Social Media Mining for Health Applications Workshop and Shared Task at AMIA, pp.1–5, 2017.

- **Sifei Han**, and Ramakanth Kavuluru."Exploratory analysis of marketing and non-marketing e-cigarette themes on Twitter." International Conference on Social Informatics, LNCS 10047, pp. 307–322, 2016.

- **Sifei Han**, and Ramakanth Kavuluru. "On assessing the sentiment of general tweets." Canadian Conference on Artificial Intelligence, LNCS 9091, pp. 181–195, 2015.

- Sarker, Abeed, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, **Sifei Han** et al. "Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task." Journal of the American Medical Informatics Association, 25 (10), 1274–1283, 2018.

- Ramakanth Kavuluru, **Sifei Han**, and Daniel Harris. "Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques." Canadian Conference on Artificial Intelligence, LNCS 7884, pp. 77–88, 2013.

- Ramakanth Kavuluru, **Sifei Han**, and Ellen J. Hahn. "On the popularity of the USB flash drive-shaped electronic cigarette Juul." Tobacco Control, 28(1): 110–112, 2019.

# Chapter 2 Related Work and Background

Social media encourages people to share their opinions and the details of their personal lives. A single post may not be meaningful health insights wise, but millions of posts with the same topic reflect a quantitative change that creates a qualitative change. Several studies have shown that aggregating millions of posts can give insights on public health. Some significant public health surveillance examples include influenza detection (Aramaki et al., 2011) and infectious disease outbreaks (Choi et al., 2016). In (Brownstein et al., 2009) it is estimated 37%-52% of Americans seek health-related information on the Internet. Usually, inaccurate or irrelevant information is also available to the public, and it is crucial to identify which information is correct, especially when it comes to health-related information. Inaccurate information could potentially have a negative impact on our well-being.

Nowadays, people prefer to use the Internet as a priority option to seek advice regarding their illnesses, drug use, and self-treatment. Chung studied the accuracy of online information regarding the safety of infants during sleep (Chung et al., 2012). The American Academy of Pediatrics has published recommendations for reducing the risk of sudden infant death syndrome (SIDS), suffocation, strangulation, entrapment, and other accidental sleep-related infant deaths. However, these recommendations are given as guidelines by health professionals containing medical jargon that cannot be easily understood by the general public without a related background in medicine. Therefore, people probably enter the keywords related to infant sleep safety into a search engine and may follow the suggestions listed in the search results.

Chuang et al. (Chung et al., 2012) analyzed 1300 websites on infant safety sleep (13 keywords and first 100 websites for each). The overall proportion of accurate information is 43.5%, inaccurate information is 28.2%, and 28.4% irrelevant information. They also found that different data sources have huge differences. Government websites (.gov or .state) and organizational websites (.org) achieved the highest level of accuracy: 80.9% and 72.5%, respectively. On the contrary, blogs and personal websites had very low accuracy score: 25.7% and 30.3% respectively. Another finding was that different keywords brought different outcomes from as high as 82% accuracy to as low as 18% accuracy. This study shows that the Internet does provide an opportunity for people to seek health-related information, and patients need skills to identify what information is most accurate. On the other hand, the study also shows the quality of keywords is important for health-related information. We can see online health information still has a long way to go to improve the accuracy of health information. It requires collaboration between health professionals, researchers, and Internet users.

Besides static websites, social network sites also contribute a massive amount of health information. The study by Chou et al. (Chou et al., 2009) shows that in 2007, about 69% of American adults had Internet access. Among Internet users, 5% of them joined in an online support group. As health providers mainly focus on clinical outcomes, patients' mental issues usually aren't given enough attention. Through affliction of emotional and physical pain, patients may develop depression and suicidal thoughts. There are several studies on suicide prevention (Kavuluru et al., 2016; Luxton et al., 2012), many identifying the most important posts during the conversion which led to a sentiment shift. By analyzing millions of posts, we may find some patterns which can help health providers and patients' families help patients through difficult times.

Monitoring sentiment change during a conversation and manual analysis is time-consuming and unrealistic. With the power of NLP and machine learning, sentiment analysis (also known as opinion mining) is used to classify the polarity of a given context automatically. More details on sentiment analysis will be discussed in section 2.1. By classifying the sentiment of posts during online communication, we can use topic modeling/statistical analysis to summarize and categorize what types of information patients will need for support, what types of support are most helpful (sharing personal stories, general support, information support, etc.), and how to attract patients and keep them in the conversation.

The benefits of online health support forums such as Cancer Survivors Network (CSN), Lungevity, and PatientsLikeMe are immense. Although the numbers of users are far below the numbers of more general social networks such as Facebook and Twitter, online health support forums offer patients the chance to interact with others who have been diagnosed with the same diseases such as lung or breast cancer. As online health support forum members are patients, health providers, patients' families and friends, the posts made by these users are more accurate than blogs and personal websites; when some posts are recognized as inaccurate, other users will quickly move to correct them.

Online health-related social media offers an abundance of information for patients, health providers, and researchers. Wicks et al. (Wicks et al., 2010) show that over 70% of PatientsLikeMe users think the site is "moderately" or "very helpful"; over 50% of patients found the site helpful for understanding the side effects of their treatments;

and 42% of patients agreed that site had helped them to find another patient who can help them understand a specific treatment for their symptoms. This shows an opportunity that online health information and communication can provide a critical mass of useful information for different parties.

## 2.1 Sentiment Analysis

### 2.1.1 Introduction

Sentiment analysis (or opinion mining) has gained significant attention from the computer science research community over the last decade due to the rapid growth in e-commerce and the practice of consumers writing online reviews for products and services they have used. Movies, restaurants, hotels, and recently even hospitals and physicians are being reviewed online. Manually aggregating all the information available in a large number of textual reviews is impractical. However, discovering different aspects of the product/service that the review is discussing and the corresponding evaluative nature of the review for each of them is computationally challenging given the idiosyncratic and informal nature of customer reviews. Sentiment analysis has also been essential in gleaning information from customer surveys that companies routinely conduct. Due to our direct involvement in an ongoing project, we also observe that companies consult researchers to conduct sentiment analysis of emails of their employees to assess personnel morale and to improve organizational behavior and decision-making. Recently, in the field of healthcare, researchers have focused on identifying emotions in suicide notes (Pestian et al., 2012) and predicting county level heart disease mortality using Twitter language usage (Eichstaedt et al., 2015).

### 2.1.2 Challenges in sentiment analysis

#### Short informal textual message

Due to the 280 character (previously 140) limit per message, Twitter user post short informal textual messages called "tweets." This type of message has brought a new challenge to sentiment analysis. They are limited in length, have many spelling errors, slang terms, emoticons (combination of numbers, letters and punctuation marks to represent facial expressions) and elongated words. They also have special markers such as user mentions that are used when users refer to each other and hashtags, which are used to search and indicate a topic or sentiment.

### Structure of language

There are three formats to express the sentiments(Hussein, 2016):

- Structured Sentiments: formal, written language to express a writer's sentiment on targeted issues.
- Semi-structured Sentiments: comes with pros and cons and is listed separately with short phrases.
- Unstructured Sentiments: the writer uses a free text format without following any constraints and sometimes implicitly to share their opinions.

Twitter users employ emoticons, slang terms and a context-free format to express their opinions. This unstructured sentiment format is very challenging for a computer to identify the sentiment polarities.

### Named entity recognition (NER)

People sharing their opinions usually target an entity/aspect; therefore, to identify the entity is important. For example, in a sentence like "I like turkey.", does the word "turkey" represent the country or the bird? Another example, "I like my new laptop, but I need a larger hard drive." The overall sentiment is positive, but the hard drive space does not meet their expectations. This is very important for aspect-based sentiment analysis.

## 2.2 Evaluation Measures

The goals of this research involve targeted sentiment classifiers to reach state-of-the-art results with limited training data; building a stance detector to understand people's opinions behind their posts better, and to improve the performance of mental health classification, and adverse drug reactions and intakes.

In the balanced dataset setting, accuracy (the proportion of all messages correctly classified over all messages) has been widely used. For our tasks are facing an imbalanced dataset, we also assess the average of the F-score of the positive and negative classes, which we will term as F-Sent for the rest of this thesis, for simplicity. We defin precision $P$, recall $R$, and F-score as

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{2.1}$$

and

$$F - score = \frac{2PR}{P+R}, \tag{2.2}$$

where *TP*, *FP* and *FN* are numbers of true positives, false positives, and false negatives, respectively. Given $\mathrm{F}_+$ and $F_-$ are F-scores for the positive and negative classes respectively, then

$$F - senti = (F_+ + F_-)/2 \tag{2.3}$$

This measure takes into account the FPs and FNs caused (including those due to neutral tweets) in classifying positive and negative sentiment categories but does not directly incorporate credit for correctly classifying neutral tweets. It is well known and has been used as the main measure in the SemEval (Nakov et al., 2013) Twitter sentiment analysis tasks. Similarly, we can calculate the macro-averaged F1 score by average the F1 of each class.

Table 2.1: Disagreement matrix to compute Cohen's kappa

| | | Annotator 1 | |
|---|---|---|---|
| | | Yes | No |
| Annotator 2 | Yes | a | b |
| | No | c | d |

In curating our own ground truth dataset, we typically ask two annotators to label the data and use Cohen's Kappa score to measure the agreement between two annotators, with equation defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.4}$$

where $p_o$ is relative observed agreement among raters, and $p_e$ is the hypothetical probability of chance agreement defined as

$$p_o = \frac{a+b}{a+b+c+d}, \; p_e = \frac{marginal_a + amrgianl_b}{a+b+c+d} \tag{2.5}$$

$$marginal_a = \frac{(a+b)(a+c)}{a+b+c+d}, \; margianl_b = \frac{(c+d)(b+d)}{a+b+c+d} \tag{2.6}$$

## 2.3 Method for Text Classification

In this section, we describe the general methods that we will use in the proposed work. It starts with a brief review of the traditional machine learning approach and feature engineering, followed by deep learning approaches. Specific methods used for particular tasks will be described in subsequent chapters.

### 2.3.1 Traditional machine learning

Traditional ML methods include naive Bayes, decision trees, $k$-nearest neighbors, logistic regression, and support vector machines(SVM). These algorithms are based on feature engineering where a set of discriminative handcrafted features is constructed to improve model performance. Although deep learning has increased in popularity recently, in most Kaggle[1] competitions currently, competitors are still winning by traditional methods. Especially when data is structured, a human can find good feature representations to train ML models. On the contrary, deep learning is adept at finding features from unstructured data such as images, audio, video. Such models can extract the features that humans cannot easily understand, but are still meaningful to a machine. For a dataset with a few hundred to a couple of thousand training samples, traditional machine learning usually outperforms deep learning methods. Since a deep learning structure is more complex than a traditional method, it has larger parameter spaces it needs to search through and learn. A small dataset cannot fully tune parameters that are generally representative for a domain leading to poor generalization. Usually, the challenge in traditional machine learning is identifying an appropriate model and features while in deep learning it is to search for appropriate architectures.

When we train a classifier, we want to have a feature vector $\boldsymbol{x} = (x_1, ..., x_n)$ to represent a training sample with a corresponding label $y \in \{0, .., k\}$, where $k$ is the number of classes. In text classification a basic feature extractor uses n-grams which is essentially the counts of word(s) and adjacent word sequences of length $n$. The n-grams are often used in a baseline model that is compared with more sophisticated models. Based on the goal of the task, we manually create features which are meaningful to the human being. In the sentiment classification, we use sentiment lexicons to get positive and negative word(s) in the posts, calculate sentiment scores, handle negation, among other things. In adverse drug reactions and medication intake message tasks, we need to use drug names to check if the drug was mentioned in the post. As we can see, the handcrafted features need to represent how humans attempt to classify the data and identify what kind of information is useful to humans to classify; these features may potentially help the machine to learn how to perform its classification. Since the machine is not human, the handcrafted features might induce FPs

[1] https://www.kaggle.com/

and FNs that need to be handled careful. Therefore, the best feature set is usually a subset of the full feature set we typically create upfront for any task; to find the best feature combination is a time-consuming work.

### 2.3.2 Deep neural networks

In the sixties, a single layer neural network was introduced and was called a perceptron. The structure was one input layer, one hidden layer, and one output layer. During that time, the machines were simple and could not handle complex operations, until the 1980s, when the multilayer perceptron (MLP) was created by Rumelhart, Williams, and Hinton. In the figure 2.1 [2] we can see that all the layers are fully connected, where each node in one layer is connected to all nodes in neighboring layers. This brings an issue when the network gets larger: the number of parameters increases drastically. For instance, with 1000 cells in the hidden layer connected with $1000 \times 1000$ elements in the input matrix, we need $10^9$ weight parameters and 1000 bias parameters for a single hidden layer. Additional layers will further increase the parameter set size. This limited the size of each layer and the depth of the network due to computing constraints. In addition, this may also lead to overfitting and the network getting stuck in local maxima.

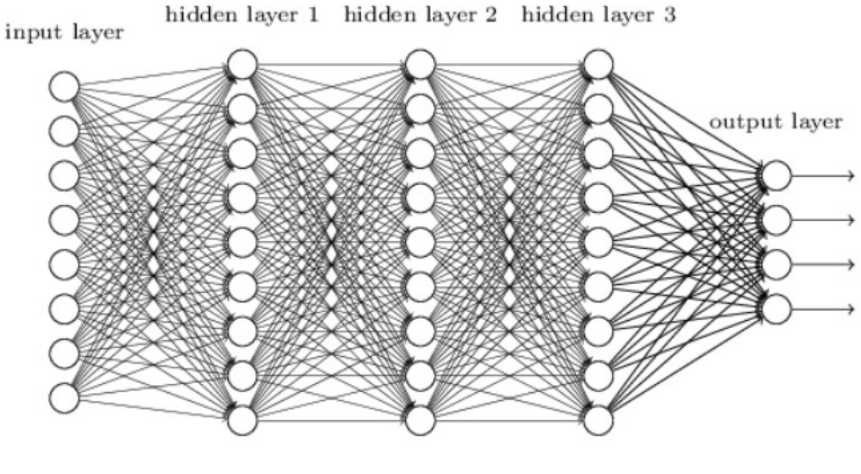


Figure 2.1: A feed forward deep neural network

To address these issues, different architectures were introduced to change the cells in the neural network such as Convolution Neural Networks (), Recurrent Neural Networks(), Long Short Term Memory Networks (), and Gated Recurrent Units().

[2] http://www.coldvision.io/wp-content/uploads/2016/07/dnn_ann_vs_dnn.png

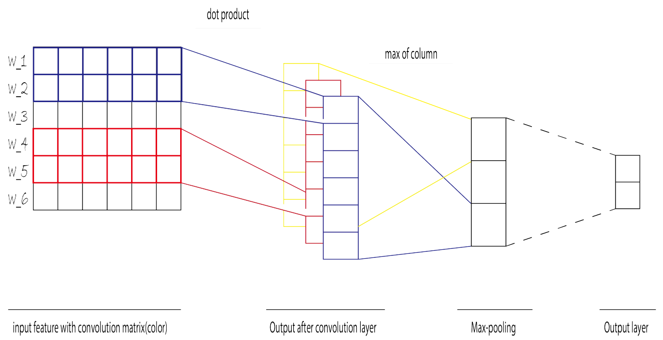

dot product
max of column
W_1
W_2
W_3
W_4
W_5
W_6
input feature with convolution matrix(color)
Output after convolution layer
Max-pooling
Output layer

Figure 2.2: The CNN model with a binary output layer for text classification

### 2.3.3 Convolution neural network

An example CNN is as shown in Figure 2.2 for the task of text classification. The difference in this network is that convolution cells only connect with a part of the input cells. It extracts local information from the connected cells in the previous layer. For instance, in the image classification task, the first hidden layer can extract some curves by looking at a part of the pixel matrix, the second hidden layer might know the combination of curves and recognize them as part of the information of an object, and the third layer might know what the object could be and give a probability for each possible class.

The purpose of an activation function is making output result from the convolutional layers compressed to a fixed real range so that the range of values that are input to the next layer is controllable; it also introduces non-linearity to extend network abilities to capture more complex functions. The common activation functions include sigmoid, ReLU, Leaky ReLU, Maxout, tanh.

A convolutional layer connected to a part of a previous layer of cells compares with a basic neural network cell. Comparing with a fully connected network (DNN), CNN reduces a lot of the parameters needed to be stored. To further lessen the size of the parameters, a pooling process can be employed. The pooling cell receives output from the previous layer and decides if this value can pass to the next layer or not. We can imagine zooming in on the image and choosing to save some meaningful pixels. This can reduce the risk of outfitting.

CNN share the weight parameters in each layer and reduce the amount of calculation required. Handcrafted features are thus not required in CNNs, as the intermediate layer outputs based on corresponding network weights can represent the features.

The deeper the network created, the more abstract information a CNN might synthesize, which generally implies a better result. The downside of CNNs is that they need a large amount of training data to tune the parameters, which requires hardware like GPUs to satisfy.

### 2.3.4 Neural attention mechanisms

The first introduction of the attention-based models in NLP was by (Bahdanau et al., 2014). The motivation was rooted in the past, where a whole sentence was encoded to a fixed-length vector; intuitively, this approach handles long sentences poorly. On the other hand, when humans perform translation, we focus on meaningful words in the sentence. The fundamental idea is that "the decoder decides what parts of the source sentence to pay attention to. By letting the decoder have an attention mechanism, we relieve the encoder from the burden of having to encode all information in the source sentence into a fixed-length vector. With this new approach, the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly. (Bahdanau et al., 2014)"

We propose to apply an attention layer after word embedding and each convolutional layer. The idea is that not all words have equal weight in signalling the classifier which labels it should assign; by adding attention weight, we want to focus on sentiment words for sentiment classification and drug names for adverse drug reaction classification. With convolution cells collecting information on the phrase level, it might provide a way to handle negation phrases (for example: not happy).

We are given a sentence $(w_1, w_2, ..., w_n)$, such that $w_i$ is the $i$-th word in it. After word embedding, we have a matrix $\boldsymbol{X}$ that contains word vectors $(\boldsymbol{x_1, x_2, ..., x_n})$. We want to add more attention to the meaningful word(s) and define the meaningful word(s) based on the task. The attention weights for the embedding layer are calculated as shown in equation (2.7) $a_i$ is the attention weight for word $i$.

$$a_i = 1 + \begin{cases} |\frac{w_i\ score}{max(lexicon\ score)}| & w_i \in lexicon \\ 0 & otherwise \end{cases} \tag{2.7}$$

This equation 2.8 shows the attention weight calculated and applied after the convolutional layer. $e_i$ is the output from a convolutional cell, $\boldsymbol{v}$ is a attention vector with random initialization. $\boldsymbol{W}$ is a convolutional filter matrix, and $\boldsymbol{a_i}$ is attention matrix.

$$e_i = \boldsymbol{v}^T tanh(\boldsymbol{W}\boldsymbol{x}_i), \quad a_i = \frac{exp(e_i)}{\sum_{k=1}^{L} exp(e_k)} \tag{2.8}$$

## 2.4 Notation

In this dissertation, we used bold upper case letters (e.g., $\mathbf{W}$) to represent matrices and vectors are represented as bold lower case letters (e.g., $\mathbf{x}$, $\mathbf{w}$). Subscripts are used to distinguish different matrices and vectors as in $\mathbf{W}_i$ and $\mathbf{x}_1$. Lower case letters are used to represent scalars (e.g., $y$).

# Chapter 3 Exploratory Analysis of Marketing and Non-marketing E-cigarette Themes on Twitter

Electronic cigarettes (e-cigs) are an emerging smoke-free tobacco product introduced in the US in 2007. An e-cig essentially consists of a battery that heats up liquid nicotine available in a cartridge into a vapor that is inhaled by the user (Etter et al., 2011), an activity often referred to as *vaping*. The broad topic of e-cig use has become a major fault line among clinical, behavioral, and policy researchers who work on tobacco products. There are arguments on either side given their reduced harm aspect ((McNeill et al., 2015) claims they are 95% less harmful than combustible cigarettes) may help addicted smokers quit smoking (Levy et al., 2016) while the long term effects of e-cigs are not yet thoroughly understood. However, there is recent evidence that vaping is linked to suppression of genes associated with regulating immune responses (Martin et al., 2016). Furthermore, based on recent news releases from the Centers for Disease Control (CDC) (Singh et al., 2016), there is an alarming 900% increase in e-cig use from 2011 to 2015 by middle and high school students who might be acquiring nicotine dependence albeit through the new e-cig product. There is also recent evidence that never smoking high school students are at increased risk of moving from vaping to smoking (Barrington-Trimis et al., 2016). In light of these findings, the FDA has recently introduced a final deeming rule (Food and Drug Administration, HHS et al., 2016) that went into effect on 8/8/2016 when regulations were extended to many electronic nicotine delivery systems including e-cigs. In this context, surveillance of online messages on e-cigs is important both to monitor the spread of false/incomplete information (Klein et al., 2016) about them and to gauge prevalence of any adverse events related to their use (Chen et al., 2013; Rudy and Durmowicz, 2016) as disclosed online.

For an emerging product like e-cigs, the follower-friend connections and "hashtag" functionality offer a convenient way for Twitter users (or "tweeters") to propagate information and facilitate discussion. An official quote we obtained earlier this year from Twitter Inc. indicates there are over 30 million public tweets on e-cigs since 2010. In our prior effort (Kavuluru and Sabbir, 2016), we found that there is a 25 fold increase in e-cig tweets from 2011 to 2015 indicating the popularity of e-cig messages on Twitter. A major amount of chatter on e-cigs on Twitter surrounds their marketing by vendors making it generally difficult to analyze regular e-cig tweets that are not dominated by marketing noise. As such, building and using a classifier that separates marketing tweets is an important pre-processing step in several efforts. We are aware of at least four such efforts (Cole-Lewis et al., 2015b; Godea et al., 2015; Huang et al., 2014; Kim et al., 2015) on building automatic classifiers for e-cig marketing tweets for various end-goals. Other researchers who studied e-cig tweets focused on sentiment analysis (Godea et al., 2015; Myslin et al., 2013) and diffusion of messages from e-cig brands on Twitter (Chu et al., 2015). In our current effort

1. We manually estimated the proportion of marketing and non-marketing tweets to be 48.6% (45.5–51.7%) : 51.4% (48.3–54.5%) from a sample of 1000 randomly selected tweets selected from over one million e-cig tweets collected through Twitter streaming API between 4/2015 and 6/2016 (Section 3.1). The ranges in parentheses show 95% confidence intervals of the proportions calculated using Wilson score (Wilson, 1927).

2. We built a classifier that achieves an accuracy of 88% (Section 3.2) in identifying marketing and non-marketing tweets using a variety of approaches ranging from traditional linear text classifiers to recent advances in classification with convolutional neural networks based on word embeddings (Kim, 2014a). Prior efforts (Huang et al., 2014; Kim et al., 2015) that seem to report similar or slightly superior ($< 2.5\%$) results estimate the proportion of marketing tweets in the dataset to be 80%–90%[1], which we find unrealistic in the current situation (based on our own assessment mentioned earlier) as public awareness and their participation in the conversation have increased.

3. After applying the binary classifier to over a million e-cig tweets, we conducted a rudimentary analysis of differences in content and user traits in both subsets (Section 3.3). We then ran topic modeling algorithms tailored for short texts (Cheng et al., 2014) on the two separate subsets by determining the ideal numbers of topics using average topic coherence scores (O'callaghan et al., 2015). We manually examined the topics generated to identify themes in general and also based on subsets of geotagged tweets at popular places of interest as identified through the GeoNames geographical database (`http://www.geonames.org`). Although prior efforts identified broad themes through manual analyses (Cole-Lewis et al., 2015a), we believe our current effort is the first to employ topic modeling to discover more specific e-cig themes (Section 3.4). Thus, rather than having investigators prede-

[1] Although achieving high F-scores for the minority class is generally difficult in heavily skewed datasets, they typically lend themselves to building classifiers with high overall accuracy across all classes or high F-score for the majority class.

termine which themes to look for in the dataset, our approach lets the dataset determine the prominent themes.

### 3.1 Dataset and Annotation

We used the Twitter streaming API to collect e-cig related tweets based on following key terms: `electronic-cigarette`, `e-cig`, `e-cigarette`, `e-juice`, `e-liquid`, `vape` and `vaping`. Variants of these terms with spaces instead of hyphens or those without the hyphens (for matching hashtags) were also used. A total of 1,166,494 tweets were obtained through the API calls from 4/2015 to 6/2016. From this dataset we randomly chose 1000 tweets to manually annotate them as marketing or non-marketing. For our purposes, marketing tweets are those that

- promote e-cig sales (coupons, free trials, offers),
- advertise new e-cig products (liquid nicotine or vaping devices), or
- review different flavors or vaping devices aiming to sell.

We (both authors) independently annotated the 1000 tweets. The labels matched for 87.3% of the tweets with an inter-annotator agreement of $\kappa = 0.726$, indicating substantial agreement (Landis and Koch, 1977). Conflicts for the 127 tweets where we chose different labels were resolved based on a subsequent face to face discussion resulting in a consolidated labeled dataset of 1000 tweets. Disagreements occurred when the marketing/advertising intent is not explicit or clear. For example, a simple message that encourages the followers to also follow the tweeter's Instagram account is not explicitly promoting e-cigs in and of itself but is nevertheless aimed toward marketing. Conflicts also occurred with reviews/recommendations when it was not clear whether a user is genuinely recommending a particular flavor that he/she has tried or whether it is the message from a manufacturer simply drawing followers' attention to their product line. While the former is not a marketing tweet, the latter would definitely fit our notion of such a message. Our final consolidated dataset has 486 marketing and 514 non-marketing tweets.

### 3.2 Marketing Tweet Classifier

The measure of performance used in this effort is accuracy, which is essentially the proportion of correctly classified tweets. We did not use the popular F-measure given we wanted to give equal importance to both classes given our aim is to study themes in both subsets of tweets. We first used linear classifiers such as support vector machines (SVM) and logistic regression (LR) classifiers as made available in the scikit-learn (Pedregosa et al., 2011a) machine learning framework. Tweet text was first preprocessed to replace all hyperlinks with the token URL and user mentions with the token TARGET. This is to minimize sparsity of very specific tokens having to do with links and user mentions and is in line with other efforts (Agarwal et al., 2011). Besides uni/bi-grams we also used as features, counts of emoticons, hashtags, URLs, user mentions, sentiment words (positive/negative), and different parts of speech in the tweet. These additional features were useful in our prior efforts in tweet sentiment analysis (Han and Kavuluru, 2015) and spotting e-cig proponents (Kavuluru and Sabbir, 2016) on Twitter. However, in this effort, considering average accuracy over hundred distinct 80%-20% train-test splits of the dataset, we did not observe any improvements with these additional features. So our final mean and 95% confidence intervals for accuracies are $88.10 \pm 0.40$ with LR and $87.14 \pm 0.45$ with SVM.

Recent advances in deep learning approaches specifically convolutional neural networks (CNNs) have shown promise for text classification (Kim, 2014a). Given our own positive experiences in replicating those approaches for biomedical text classification (Rios and Kavuluru, 2015), we also applied CNNs with word embeddings to generate feature maps for marketing tweet classification. The main notion in CNNs is of so called *convolution filters* (CFs) that are traditionally used in signal processing. The general idea is to learn several CFs which are able to extract useful features from a document for the specific classification task based on the training dataset. In the training phase, the inputs to the CNN are projections of constituent word vectors (which are typically randomly initialized) from a fixed size sliding window over the document. Model parameters to learn include the word vectors, the convolution filters (which are typically modeled as matrices), and the connection weights from the convolved intermediate output to the two nodes (for binary classification) in the output layer. Due to the nature of this particular paper, we refer the readers to our recent paper (Rios and Kavuluru, 2015, Section 3) for a detailed description of CNN models including specifics of parameter initialization and drop-out regularization (to prevent overfitting). Averaging the $[0, 1]$ probability estimates of the corresponding classes from several (typically ten) CNNs seems to help in getting a more robust model. We ran ten such models (each with ten CNNs, so a total of 100 CNNs) on ten different 80%-20% train-test splits of the dataset. The corresponding accuracies were: 89, 88.5, 85.5, 86, 87, 90.5, 87.2, 88.5, 90.5, and 89 with an average of 88.17%, which is only slightly better than the mean accuracy obtained using logistic regression.

### 3.3 Characteristics of Marketing/Non-Marketing Tweets

As discussed earlier, although the ability to separate marketing tweets from those that do not have that agenda is of interest in and of itself, in this effort, we wanted to study themes evolving from both subsets of the dataset. We applied all three classifiers (SVM, LR, and CNN) built in Section 3.2 using all hand-labeled tweets to all 1,166,494 tweets in our full dataset. We considered those tweets for which all three classifiers predicted the same label, which turned out to be for 1,021,561 (87.56% of the full dataset) of which 456,290 (44.66%) were predicted to be marketing and 565,271 (55.34%) belonged to the other class. To get a basic idea of the tweet content, we simply counted and sorted the words in each subset in descending count values. The top 20 words in both subsets are

- *Marketing*: win, vaporizer, free, mod, get, enter, giveaway, new, premium, code, shipping, bottles, USA, use, box, promo, kit, available, follow, DNA
- *Non-Marketing*: smoking, new, use, rips, like, cigarettes, via, man, get, tobacco, health, video, study, FDA, ban, one, smoke, people, news, explodes

Even with this simple exercise, we notice that the marketing tweets are dominated by e-cig promotions and sales terms or devices for vaping (mod, vaporizer, kit). On the other hand, terms in the non-marketing tweets are about tobacco smoking, health studies, and FDA regulations.

Table 3.1: Content and user characteristics of the datasets

| | Marketing | Non-marketing |
|---|---:|---:|
| E-cig flavors | 25472 | 4612 |
| Harm reduction | 19 | 2256 |
| Smokefree aspect | 553 | 3201 |
| Smoking cessation | 6363 | 22421 |
| Contain “FDA” | 204 | 18297 |
| Number of unique users | 66,957 | 231,982 |
| User handles containing e-cig terms | 4777 (7.1%) | 3859 (1.7%) |
| Avg. # tweets per user | 6.81 ($\sigma = 197$) | 2.44 ($\sigma = 84$) |

Next, we look at specific content and user characteristics of both subsets. In our prior work (Kavuluru and Sabbir, 2016), we analyzed the tweets generated by e-cig proponents tweeters along four well known broad themes. We developed regular expressions (please see (Kavuluru and Sabbir, 2016, Section 5.3)) in consultation with a tobacco researcher to capture tweets belonging to these themes. As part of the preliminary analysis, in this effort, we applied those regexes to the two subsets of tweets and obtained the corresponding numbers of thematic tweets shown in the first four rows of Table 3.1. Except for e-cig flavors, which are a well known major selling point, the non-marketing datasets contain more tweets in the three other themes (even after accounting for the slight variation in dataset sizes). It is still disconcerting to see the 6363 (1.4%) marketing tweets discussing smoking cessation when long term consequences of e-cig use are still being investigated. We also looked at how many tweets mention FDA and as expected the majority belong to the non-marketing class.

The last three rows of Table 3.1 deal with user characteristics of both datasets. We notice that there are 3.5 times as many unique tweeters in the non-marketing set as in the marketing class (row 6). We clarify that some users can belong to both the marketing and non-marketing class if they generate tweets in both datasets. In fact, the top non-marketing tweeter `@ecigitesztek` has 37,949 such tweets but is also ranked 2nd among tweeters in the marketing group with 27,019 tweets. A cursory examination of this public profile indicates that it belongs to a Hungarian vaping aficionado who almost exclusively tweets about e-cigs and at the time of this writing (re)tweeted over 153,000 times. However, with 11,186 tweeters common to both datasets corresponding to counts from row 6, the Jaccard similarity coefficient is only 0.03. Given marketers tend to use appealing user handles that indicate their purpose, we counted the number of user handles that contain e-cig popular terms such as ecig, vapor, vapour, vape, vaping, eliquid, ejuice, and smoke as substrings of the user handle. 15 out of the top 20 tweeters in both datasets contain one of these terms as a substring. From row 7, we see that more than 7% of the marketing profiles satisfy this compared with only 1.7% from the other class.

The final row indicates the average number of tweets per user with standard deviations in parentheses; the difference in the averages is not surprising but the standard deviation magnitude in the marketing set being more than twice that in the other class is revealing in that few users are responsible for many marketing tweets. To further examine this phenomenon, we plotted the cumulative proportion of tweets in the corresponding datasets contributed by the top $10, \ldots, 100$ tweeters in Figure 3.1. It is straightforward to see that the top tweeters in the marketing dataset generate twice the proportion of tweets as generated by the corresponding top users in the non-marketing dataset. Although the Jaccard coefficient between tweeter sets

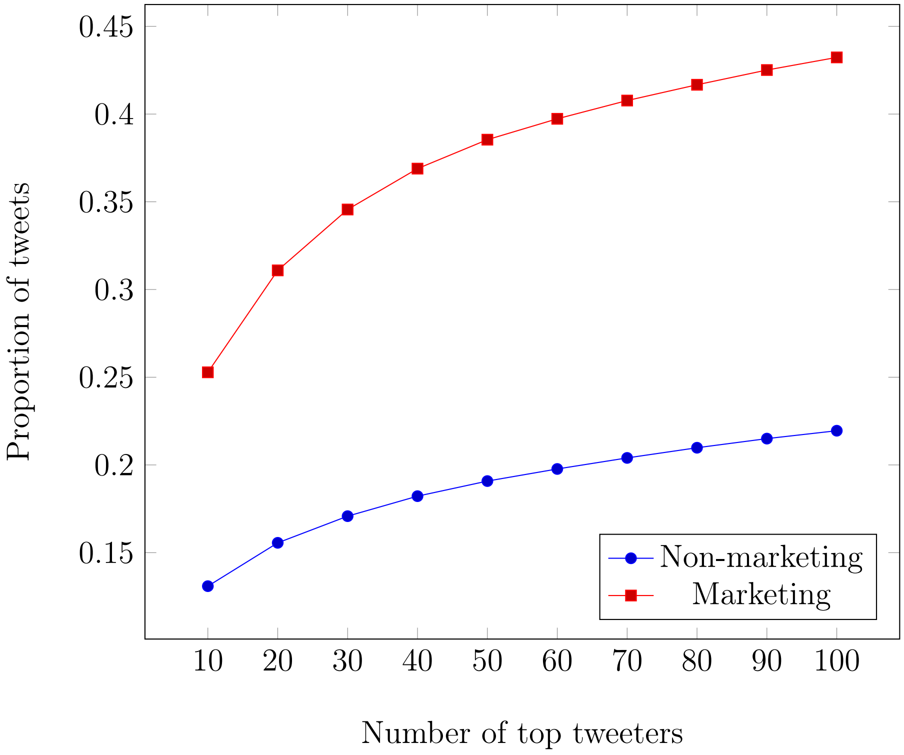


Figure 3.1: Proportion of tweets via top $10, \ldots, 100$ tweeters

from both datasets in only 0.03, when considering only top 100 tweeters from both datasets, 84 of the top 100 marketing tweeters have generated non-marketing tweets; 88 of the top 100 non-marketing tweeters also authored marketing tweets.

### 3.4 Themes in Marketing/Non-Marketing Tweets

To dig more into these two subsets of tweets, we applied the Biterm Topic Modeling (BTM) (Cheng et al., 2014) approach, which is specifically designed for short text messages like tweets, to these marketing and non-marketing tweets subsets separately. Given recent results that demonstrate that aggregating short text messages such as tweets can lead to better modeling (Hong and Davison, 2010), we partitioned the datasets into groups of ten tweets each where each such group is treated as a short document before applying BTM. Besides using the same tweet pre-processing techniques used for classification, we additionally removed commonly occurring terms from the tweets such as stop words and frequent terms such as the key words used to search for e-cig tweets (e.g., e-cig, vape, vaping, vapor, eliquid) given we already know the tweets are on the general topic of e-cigs.

### 3.4.1 Topic modeling configuration

Most topic modeling approaches have the inherent requirement that the user suggest the number of topics $k$ to fit to the corpus. It is often tricky to pick a specific $k$, which is generally chosen by trial and error based on human examination of topics generated with different settings of $k$. We circumvented this potentially tedious and subjective exercise by using a recently introduced measure of topic coherence by O'Callaghan et al. (O'callaghan et al., 2015) based on neural word embeddings. Topic coherence is a direct measure of intrinsic quality of a topic. For each topic $T$ generated, let $w_1^T, \ldots, w_N^T$ be the set of top $N$ words according to the $P(w|T)$ distribution resulting from the topic modeling process. Then the coherence of $T$ parameterized by $N$ is

$$\mathcal{C}_N^T = \frac{1}{\binom{N}{2}} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \cos(\vec{w_i^T}, \vec{w_j^T}),$$

where $\vec{w_i^T} \in \mathbb{R}^d$ is the dense vectorial representation for the corresponding words learned through the continuous bag-of-words (CBOW) word embedding approach, which is part of the popular word2vec package (Mikolov et al., 2013). We picked dimensionality $d = 300$ and word window size of five for the CBOW configuration in word2vec and ran it on the full corpus of e-cig tweets. Given this definition of average coherence, the idea is to pick $k \in \{10, 20, 30, \ldots, L\}$ that maximizes the weighted average coherence (WAC) across $k$ topics

$$\sum_{i=1}^{k} P(T_i) \cdot \mathcal{C}_N^{T_i}, \tag{3.1}$$

where $P(T_i)$ is the probability estimate of the prominence of topic $T_i$ in the corpus (from BTM output), $N$ is the number of top few terms chosen per topic (typically 10 or 20, the latter is used in this paper), and $L$ is chosen to be 50. Note that cosine similarity measure we use here scores term pairs that are semantically similar higher than pairs of words related in a different fashion. This does not, however, affect the validity of our topic coherence approach given topics that contain highly similar words are generally more coherent and simpler to interpret than those that contain words that are related in a more associative manner.

### 3.4.2 Prominent e-cig themes

We recall that topic models output several parameters (Blei and Lafferty, 2009) including a distribution of topics per document (topic proportions: $P(T|d)$) in the corpus and also the distribution of words per topic (per-topic term probabilities: $P(w|T)$), where $T$ is a topic, $d$ is a document, and $w$ is a word. In general, a topic is visualized by displaying the top $N$ (the variable in equation 3.1) words $w$ according to $P(w|T)$. However, a human agent still needs to look at the top $N$ terms of the topic and identify/interpret a semantic *theme*. This is the distinction we use in this effort too – a topic is a group of $N$ words/terms sorted in descending order according to $P(w|T)$ and a theme is a semantic interpretation of what the topic represents based on our manual review. Even though topic modeling research has come a long way, interpretation of resulting topics for exploratory purposes involves significant manual effort, albeit guided by output distributions mentioned earlier. The rest of this paper involves such exploration to grasp the underlying themes.

Based on our experiments, we found that $k = 10$ maximizes the WAC in equation 3.1 for the marketing tweets in the corpus. The corresponding value for non-marketing tweets is $k = 50$. This is not surprising given marketing tweets are expected to contain fairly predictable themes that are favorable to e-cigs in general encouraging tweeters to buy/try them or sign-up for more offers. However, the non-marketing subset is more diverse given it is essentially a catch-all for all other topics about e-cigs. Next, we discuss some topics from both subsets.

**Marketing Themes**: Upon manual examination of the ten topics from the marketing tweets (MT) corpus, we notice a few that are clear and reflect expected themes from this subcorpus. Here we show three of those topics enumerating some of the top 20 words in the topic. The words are rearranged slightly to better reflect the theme on hand. (However, all words are still from the list of top 20 terms for the topic; otherwise, our analysis would be self-deceiving.)

**MT1**: free, shipping, code, promo, win, purchase, prizes, enter, giveaway

**MT2**: vaporizer, pen, mod, kit, battery, portable, starter, electronic, atomizer

**MT3**: premium, line, lab, certified, AEMSA, cleanliness, consistency, wholesale

The first topic represents the theme of promotional activities involved in marketing e-cigs. The second theme involves vape pens or devices that actually vaporize the liquid nicotine to be inhaled by vapers. The third topic surfaces an unexpected theme of marketing activities that also highlight the quality of the e-liquid products through independent lab certifications offered by the registered nonprofit organization American E-liquid Manufacturing Standards Association (AEMSA), which was established in 2012 for the purpose of promoting safety and standardization in manufacturing liquid nicotine products.

**Non-Marketing Themes**: The following is the list of major topics in the non-marketing tweets (NT) corpus.

**NT1**: lungs, cells, flavors, toxic, effects, exposure, study, damage, aerosols

**NT2**: FDA, poisonings, calls, surge, skyrocket, nicotine, poison, children

**NT3**: explodes, coma, teen, mouth, burns, injured, suffers, neck, hole, hospital

**NT4**: FDA, tobacco, industry, market, regulation, product, ban, deeming, rule

**NT5**: tobacco, laws, CASAA, smoke, healthier, alternative, FDA, grandfather

**NT6**: quit, smoking, help, current, smokers, cigarette, users, NHS, review

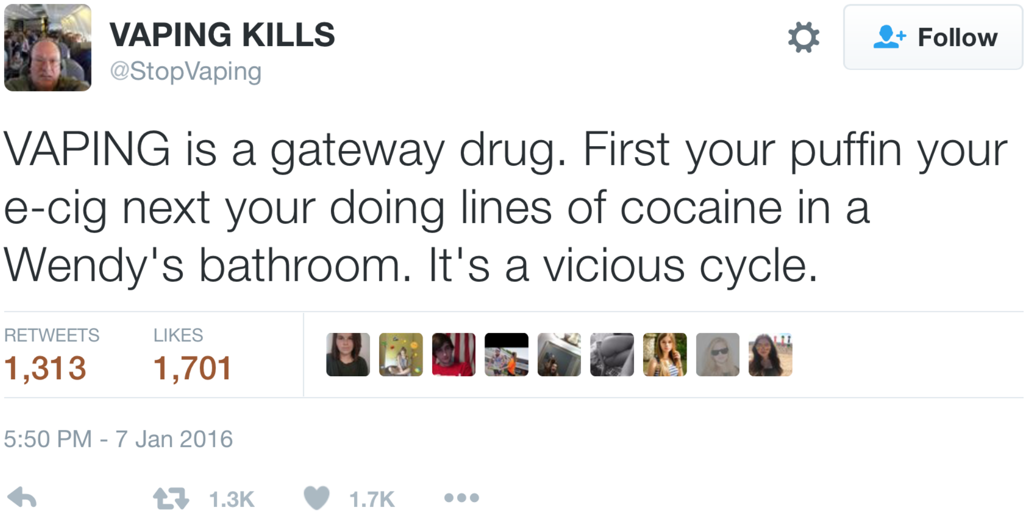
**NT7**: teen, smoking, CDC, study, middle, school, students, tripled, fell

**NT8**: ban, Wales, government, public, enclosed, spaces, pushes, ahead

**NT9**: gateway, drug, doing, cocaine, bathroom, lines, puffin, Wendy, heroin

Note that we only report nine topics here because we found these to be most interesting and also given several others seemed very similar to these nine. There are also a few that do not seem to indicate a specific non-trivial theme and hence were excluded. The first theme NT1 is about toxic effects of e-cigs. An examination of biomedical articles with the search terms `e-cigarettes` AND `toxic` AND `lungs` returned several articles discussing experiments that demonstrated how flavoring agents of e-cigs, and not the liquid nicotine itself, are responsible for toxic effects of inhaling e-cig vapors. NT2's theme relates to a news piece that diffused through Twitter about FDA receiving many calls involving poisoning complaints by e-cig users. NT3 and several related topics (not displayed here) discussed explosions of the vaping devices while in use resulting in burns and hospitalizations (Rudy and Durmowicz, 2016). NT4 represents a general theme involving FDA regulatory activities and the new deeming rule (Food and Drug Administration, HHS et al., 2016), which was thought to be impending throughout the past few years.

In NT5, we see a very specific theme that involves the non-profit organization Consumer Advocates for Smokefree Alternatives Association (CASAA) and the general harm-reduction perspective of e-cigs as an healthier alternative to cigarettes for people who want to quit smoking. The last term 'grandfather' in NT5 refers to new regulations extending to any product introduced/modified on or after the so called grandfather date set to 2/15/2007 by the FDA (Food and Drug Administration, HHS et al., 2016). This date is critical to many e-cig businesses as all those products (already in market) will now be subject to the new FDA regulations and hence need to be approved by it. NT6 represents the theme of using e-cigs as an aid to smoking cessation. The term NHS refers to UK's National Health Service, which has taken a favorable stance to e-cig use for treating addicted smokers (McNeill et al., 2015). NT7 is about research reports by the CDC indicating tripling of current e-cig use by middle and high school students from 2013 to 2014 (Centers for Disease Control, 2015). NT8 highlights another news piece on Wales (of UK) government passing a law to ban vaping in enclosed spaces.



VAPING KILLS
@StopVaping

VAPING is a gateway drug. First your puffin your e-cig next your doing lines of cocaine in a Wendy's bathroom. It's a vicious cycle.

RETWEETS 1,313
LIKES 1,701

5:50 PM - 7 Jan 2016

Figure 3.2: Tweet leading to topic NT9 on e-cigs as a gateway drug

The final topic NT9 is unusual and seems to indicate e-cigs as a gateway drug to use other more harmful products such as cigarettes, cocaine, and heroin. Although there is some evidence (Barrington-Trimis et al., 2016) to support this idea, this particular topic appeared atypical with words like bathroom, lines, and Wendy. A deeper examination revealed that most of the words in this topic are mostly coming from one tweet shown in Figure 3.2. As can be seen, this tweet was retweeted more than 1000 times. Given retweets are essentially a reasonable and natural mechanism to add more weight to a particular topic, we decided to not to delete them in our analysis. However, this particular topic led us to dig deeper into manifestations of topics of this nature. There were two other non-marketing topics like this based on frequent retweets or many tweets involving some minor modifications of a very specific tweet: one involved a picture of film actor Ben Affleck vaping after getting a traffic rule violation ticket (the topic had words Ben, Affleck, and ticket) and another involved the URL of an online petition offering support to the then UK prime minister David Cameron and other politicians trying to block certain e-cig regulations in the UK.

**Effect of Excluding Retweets**: Given this observation involving NT9, we wanted to study the effect of retweets on topic modeling. We found that 36% of marketing tweets and 43% of non-marketing tweets were due to retweets. Thus we see that retweets constitute a significant proportion of the full datasets. We generated new topic models with these subsets excluding all retweets to see if there is a noticeable difference in the themes. Although the themes did not change significantly, the words used to represent the topics have changed slightly in most cases. For example, the theme in this new set of topics corresponding to NT9 had the following top words: gateway, drug, smoking, heroin, cocaine. None of the specific words (bathroom, doing, puffin, lines, Wendy) from the highly retweeted message in Figure 3.2 showed up in the new topic. There were no other topics indicating a gateway theme. There was no topic involving Ben Affleck's traffic ticket but the petition related topic involving former UK prime minister David Cameron was apparent with slightly different words. All other themes NT1–NT8 were evident in the new set of topics. There was only one new theme that wasn't already in the topic set from the full dataset. This was mostly about vaporizer/e-liquid brand names with top terms including: sigelei, hexohm, flawless, ipvmini, districtf, tugboatrda, appletop, longislandbrewed. There was no major change in the themes for the marketing tweet subset.

Finally, we wanted to see who is tweeting on various themes identified through our approach. To this end, we picked two different non-marketing themes, NT6 (e-cigs for smoking cessation) and NT7 (CDC reporting on increasing teen vaping). For each of the corresponding topics $T$, we ranked all tweets $s$ according to $P(T|s)$. Based on the authorship of the top 10,000 tweets according this ranking, we sorted tweeters in descending order based on the counts of top 10,000 tweets they authored. We manually examined the top few ranked tweeters in this list. For theme NT7, 11 out of 20 top tweeters are regular people tweeting about e-cigs but only 2 out of 20 top tweeters for NT6 are regular tweeters; the other tweeters being institutions or companies that have a clear positive stance for and commercial interest in e-cigs. This indicates that regular tweeters (even if they are in favor of e-cigs) are more inclined to tweet about news involving e-cigs, even when it is not favorable. Commercial tweeters tend to exclusively focus on propagating favorable news pieces besides promoting their products.

Overall, our effort offers a complementary approach by surfacing specific themes in comparison to manual coding (Cole-Lewis et al., 2015a; Kavuluru and Sabbir, 2016) where only broad topics such as smoking cessation, flavors, and safety are typically used. This is our main contribution – demonstrating the feasibility of topic modeling based thematic analysis of e-cig chatter on Twitter. Some of our extracted themes may already be common knowledge for tobacco researchers who regularly follow e-cig related news. But we believe the topic modeling approach can help surface a more comprehensive set of themes with less manual exploration burden. It also gives a better sense of the strength of a theme (as observed by the the corresponding topic's ranking) and main tweeters authoring the corresponding thematic tweets.

### 3.4.3 Themes in geotagged tweets

Geotagged tweets with the associated latitude and longitude information offer a different lens to understand e-cig messages. There have been very few studies examining the locations where e-cigs are used. There is only particular study (Kim et al., 2015) that we are aware of where prepositional phrase patterns were used over tweet text to identify e-cig use in a class, school, room/bed/house, or bathroom. In our effort, we are not necessarily concerned about e-cig use, but are generally interested in knowing themes from tweets generated near different types of places of interest. Our dataset has a total of 3208 geotagged tweets which is less than 1% of the full dataset. Using the GeoNames API (`http://www.geonames.org/export/web-services.html`), we identified the nearest *toponym* for each of the corresponding geocodes using the `findNearby` method. In our dataset, the average distance between the geo-code and nearest toponym was 300 meters. Toponyms can be names of larger geographical areas such as cities or rivers, but can also refer to small locations such as a school, hospital, or a park. Each toponym (e.g., University of Kentucky) is associated with a corresponding feature code (e.g., UNIV).

We aggregated tweets based on feature codes (`http://www.geonames.org/export/codes.html`) of the toponyms returned and obtained the following distribution (top ten codes) where counts are shown in parentheses:

*hotel (596), populated-place (411), church (314), school (311), building (286),*

*mall (158), park (109), lake (91), library (80), and post office (74).*

In addition to these we also considered, travel end-points (81) as a single class (airports, bus stations, and railway stations), restaurants (39), hospitals (45), museums (13), and universities (11). A simple string search revealed that in very few cases the geotagged tweet content actually made explicit connection to the corresponding feature code. We were able to find 2–3 tweets at hotels, schools, and airports indicating the location type as part of the tweet (e.g., "vaping in class" and "flight is full"). Except for schools, parks, restaurants, hospitals, and airports, all locations had more marketing tweets than regular tweets. Overall, 52% of geotagged tweets belonged to the marketing class, a 7.5% increase compared with the corresponding proportion in the full dataset as discussed in the beginning of Section 3.3.

For each of these different location types, we identified top topics by fitting topic models to the corresponding sets of tweets. Given marketing tweets have a clear agenda, we only look at non-marketing top topics. For clarity, we simply outline the theme without listing all the keywords

- Church: Ban on e-cigs for minors in Texas
- Hotel: E-cig use rising among young people
- Park: Pros and cons of E-cig regulations
- School: Smoking rates fall as e-cig use increases among teens

Other locations either did not have a significant number of tweets or had tweets without any dominant theme. We realize that our analysis in this section may not be precise in the sense that tweets originating from different types of places may not be from people who are visiting those places for relevant purposes; tweeters might simply be around those places when they tweet. However, we believe with a large exhaustive dataset spanning multiple years, given we only look at top themes, we can arrive at themes that are representative of people visiting those places.

### 3.5 Conclusion

E-cigs continue to survive as a controversial tobacco product and are currently subject to new FDA regulations since 8/8/2016 with a grandfathering date set to 2/15/2007. The FDA, biomedical researchers, physicians, tobacco industry, and most important the nation's public are all key players whose activities will be affected with these products for the foreseeable future. Public health and tobacco researchers are split in their opinions regarding e-cig use by smokers who would otherwise continue with regular cigarettes. Computational social science and informatics approaches can offer a more objective lens through which the social media landscape of e-cigs can be gleaned for online surveillance of both product marketing practices and adverse events.

Although prior efforts exist in content analysis based on pre-determined broad themes, we do not see results on automatic extraction of themes from social media posts on e-cigs. We believe computational approaches provide an important avenue that can complement traditional survey based research efforts considering the cost and time factors involved in the latter case. Twitter in particular has been well studied in the context of public health informatics efforts and provides a major platform for e-cig chatter on the Web.

In this chapter , we conduct thematic analysis experiments involving over a million e-cig tweets collected during a 15 month period (4/2015 – 6/2016). To deal with the major presence of marketing chatter, we first built a classifier that achieved an accuracy of over 88% in identifying marketing and non-marketing tweets based on a manually labeled dataset. We conducted preliminary content and user analysis of marketing and non-marketing tweets as classified by our model. Subsequently, we fit topic models to the two subsets of tweets and interpreted them to identify specific themes that were not apparent in manual efforts. This is not surprising given the fast changing discourse on e-cigs creates a corresponding rapidly evolving social media landscape. This, however, points to an important weakness of our approach – it is not *online*, where new e-cig tweets continuously collected through the Twitter streaming API are used to generate new topics as enough evidence accumulates. As part of future work, we plan to employ online topic models (Hoffman et al., 2010) and facilitate their exploration using well known topic browsing approaches (Chaney and Blei, 2012; Malik et al., 2013). Nevertheless, here we provide what we believe is a first strong proof of concept for employing topic models to comprehend evolving e-cig themes on Twitter. Given gender, age group, race and ethnicity can be predicted with reasonable accuracy (Culotta et al., 2015; Liu and Ruths, 2013; Nguyen et al., 2013), an important future research direction is to use these methods to classify e-cig tweeters into these demographic categories and identify e-cig themes in tweets authored by specific subpopulations. For example, given african american teenagers are an active group on Twitter (Pew Research Internet Project, 2013a), identifying popular e-cig themes authored by them (including retweets and favorites) may yield insights specific to that demographic segment. Similar analysis can also be conducted with tweets originating from rural areas given the typical firehose is dominated by urban tweeters.

# Chapter 4 Message Triaging for Mental Health Forums

Mental disorders are a critical health issue among children and adolescents. According to National Alliance on Mental Illness (NAMI, 2019), 20% of youth ages 13-18 have, or will have a severe mental illness; 70% of youth in state and local juvenile justice systems have a mental illness; suicide is the third leading cause of death in youth ages 10-24, and 90% of those who died by suicide had an underlying mental illness. These statistics show we need to pay attention to teenagers' mental health, arguably more than that for their physical ailments. Given adolescents may not typically disclose mental ailments to parents and are less keen on visiting doctors, it is often tricky to identify symptoms and seek help. With the advent of social networks, online support forums provide an alternative to help teenagers who have mental issues by enabling a safe space that they can use to seek help and also belong to a community of folks going through similar issues.

As per WHO[1], the number of psychiatrists working in the mental health sector per 100,000 people ranges from 0 to 40.98 with mean 4.178 and median 1.108. This statistic is much worse in rural and low-resource areas.

In this chapter, we introduce a machine learning approach to help human mental health support forum mentors to identify users who need immediate help based on the their posts. ReachOut is an online mental health support forum which focuses on young people aged from 14 to 25. ReachOut has trained mentors to help the teenagers get through their tough times. However, each month, there are 110,000 visitors accessing the forum[2]. This is very challenging for moderators to manually prioritize which users need more attention. Teaching computers to triage the messages is one way to greatly reduce the time delay between when a crisis indicating message is posted and a human moderator gets in touch with the poster.

The rest of this chapter is organized as following: background and related works are addressed in section 4.1. Method and experiments are described in sections 4.2 and 4.3, respectively. Conclusion and future study are discussed in section 4.4.

[1] http://apps.who.int/gho/data/node.main.MHHR?lang=en

[2] https://parents.au.reachout.com/frequently-asked-questions

## 4.1 Background and Related Works

In 2016 and 2017, the Computational Linguistics and Clinical Psychology Workshop (CLPsych) held two consecutive online shared tasks with the ReachOut forum message data. The goal of the task is to create an automatic system to prioritize the mental health forum messages by how urgently they require moderator attention. The task itself is a four-class classification problem, and the classes are green, amber, red, and crisis based on the level of urgency with crisis being the most severe category. In 2016, 15 teams participated in the shared task. Three teams tied with averaged F1-score on the non-green class of 0.42. Kim et al. (Mac Kim et al., 2016) had the best model trained with three SGD classifiers whose output probability estimates were averaged to get final predictions. Malmasi et al. (Malmasi et al., 2016) trained a stacking model with 100 SVM base models with a random forest as the meta-classifier. Brew's (Brew, 2016) best model involved an SVM with RBF kernel. In 2017, there were 16 teams and the winning team used voting on a numbers of SVM classifiers. The ensemble approach created all the top performing models with base models being traditional classifiers.

In this chapter, we focus on the CLPsych 2017 shared task where the class distribution for training and testing set are as listed in table 4.1.

Table 4.1: Distribution of labels across training and testing data

| | Train | | Test | |
|---|---|---|---|---|
| | Count | % | Count | % |
| Amber | 296 | 25% | 94 | 23.5% |
| Crisis | 40 | 3.5% | 42 | 10.5% |
| Green | 715 | 60% | 216 | 54% |
| Red | 137 | 11.5% | 48 | 12% |

In addition, the organizers also provided over 150,000 unlabeled posts. The official measurement is the average F1-score on non-green classes.

## 4.2 Method

We used a supervised machine learning approach to triage the messages. In this section, we describe the pre-processing on the posts, feature extraction, feature selection, and the method. During the pre-processing step, we used BeautifulSoup and XMLtoDict python package to process the XML format files provided by the organizer. All the non-ASCII characters were removed, posts were lower cased, and user mentions were replaced by the token TARGET. We used the Gensim python package to train the word2vec model on posts with 50 dimensions and built a topic model with 20 dimensions. We created a total of 5040 features. The context-based features include n-grams, Word2Vec representations, topic model induced distributions, and Doc2Vec representations. The metadata features include the time of post, modified time, and the type of board the message posted to. Lexicon based features included those derived from NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013) and LIWC (Tausczik and Pennebaker, 2010). Author's information is also incorporated as meta features. After feature extraction, we applied feature selection based on a statistical approach: we calculated the sample standard deviation for each feature across all its values from all instances. We filtered out features whose SD is less then 0.0075. Intuitively, features whose values do not vary across instances are not as helpful given they may not be able to help distinguish between instances of different classes.

### 4.2.1 Word2Vec representation (50 features)

We used the Gensim Python library (Řehůřek and Sojka, 2010) to train the skip-gram model on all the post bodies with 50 dimensions. The parameters used in the training skip-gram model include: the minimum count for the word frequency set to five, discarded widespread words, parameter that controals downsampling frequent words set to $10^{-3}$, and negative sampling and window size both set to 15.

### 4.2.2 Doc2Vec representation (100 features)

Instead of averaging the word2vec vectors for each word in the post, we trained Doc2Vec by Gensim Python library (Řehůřek and Sojka, 2010) with the distributed memory model of paragraph vectors. The dimensions are 100, the window size is 15, and the minimum count is two.

### 4.2.3 N-grams (1000 features)

We extracted 1000 most frequent unigrams and bigrams from all post bodies and applied TF-IDF transformation.

### 4.2.4 Lexicon scores (20 features)

Words are generally known to represent certain emotions as curated by experts; therefore, we used a sentiment lexicon to identify the positive and negative emotion words. We created 11 features based on NRC-lexicon (Mohammad et al., 2013), including positive and negative sentiment scores and word counts for unigrams or bigrams, the sums of positive and negative sentiment scores of unigrams and/or bigrams. We also used LIWC (Tausczik and Pennebaker, 2010) dictionary to assess the sentiment from different degrees, including anxiety, anger, sadness, positive emotions class, negative emotions class, death, risk, swear words, and tone.

### 4.2.5 Author features (24 features)

The goal is to characterize the author of a post and use it in assessing the post's severity. We created a list of author features including a one-hot vector for type of author (the size of vector is 20), the age of the author's account in months, how frequently the author posts messages, how many times was the author mentioned before current post and a Boolean feature that sees if the author is a forum moderator or not.

### 4.2.6 Meta features (5 features)

We derived the features from post metadata including: the time, date, the month of the post; the post has been modified or not; if modified, whether it is by the author or another moderator.

### 4.2.7 History features (500 features)

People sharing their feelings in the support forum and get comments from others. Therefore, the previous messages posted by the user or the messages in the same thread can help assess the mental health of the forum user. In this history feature, we categorized them into two types: thread history and user history. In both histories, we select the five most recent posts within seven days from the current post. Each post is represented as a vector of 50 dimensions via Word2Vec.

### 4.2.8 Mention features (50 features)

We believe the interaction between users could provide hidden clues about the eventual classification. Communication can help mentors provide support to forum users.

People with similar thought processes might be reaching out to and mentioning particular forum moderators. Thus, a message that mentions a specific moderator might be thematically related to other messages that also mention this moderator. So we introduce a new feature vector of length 50 which is set to the average of all words in all messages that contain the moderator mentioned in the current message. If a message does not mention any moderator, this feature vector is set to the zero vector.

### 4.2.9 Handcrafted features

The traditional machine learning requires a good feature set and thus feature engineering plays an important role in creating a robust model. We categorize these handcraft features as following:

#### Negation (2 features)

Negations involves use of words that create the opposite meaning of the main sentiment word used in a message. For instance, *I am feeling good today.* and *I am not feeling good today.* The word *not* clearly changed the polarity of the sentence. People who need immediate support usually express strong negative sentiments and negation is used more frequently than normal people. Therefore, we count how many negation words used in the sentence and percentage of negation words in it. The negation words include *not, no, dont, cannot, cant, never* and variations form of *nt́* such as *don't, doesn't, isn't.*

#### Post context lexicon (46 features)

We also read these urgent posts and identified negative sentiment terms used in the posts. By analyzing the phrases used by the users, we selected the following terms: *kill, scare, scared, scaring, useless, hopeless, crash, depression, depressive, psychosis, harming, die, self harm, stupid, psych, hurt, anyone, anymore, anything, anyway, give up, struggle, stress, anxiety, tried, over, hard, idiot, suicide, suicidal, suffocate, want to stop, go away, self harm, unsafe, end, screw, fuck, dead, depressions, depressing, depressed, psychoses, harm, gave up, given up.* This lexicon creates a count based feature vector.

#### Negated terms (1 feature)

This is a count feature that captures number of negated terms (that is, preceded by a negation word), among the words *care, stop, safe, help.*

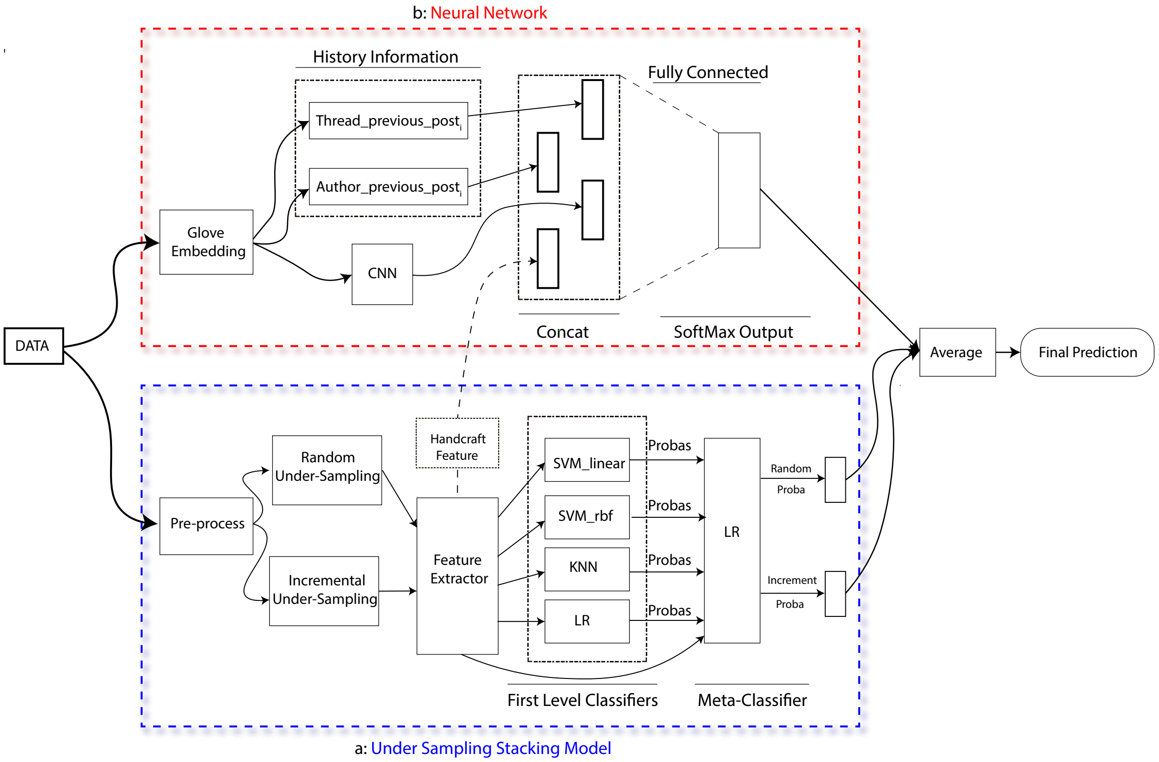


Figure 4.1: DNN with under sampling stacking architecture

### 4.2.10 Model

In this section, we start with the traditional machine learning approach followed by the deep learning approach and end with discussion of the model we used in shared task and the current best model.

#### Traditional machine learning

The overall architecture of the framework is displayed in Figure 4.1, where the bottom portion indicates the application of linear models as well as the well-known k-NN and nonlinear kernel-based approaches. The crucial step involves undersampling (Wallace et al., 2011) the training data of more frequent classes to deal with class imbalances. Given the crisis class is the smallest but is also the most important for triaging purposes, a bagged classifier created with the undersampling technique applied to all other classes. Fifty meta-models were bagged (via model averaging), where each has a stacked structure composed of four models: logistic regression (LR), support vector machines (SVM), SVMs with the RBF kernel, and the k-NN model. Undersampling is done in two ways for the meta-models: (1). In the random undersampling approach, for each model C instances were randomly selected from all other classes where C is the number of crisis examples in the training dataset. (2). In the incremental undersampling approach, the number of instances from the other classes are gradually increased, starting with only C instances, and then to reach the full instance sets for all classes in the 50th meta-model. In both these approaches, the number of instances for the crisis class is always constant and equal to its training dataset size. The outputs from the bagged meta-models are averaged to generate two $[0, 1]$ scores one for each undersampling method. These two scores are subsequently averaged with the probability estimate from the DNN ensemble. Features are derived from the contents of three different text sources: (1). The original post to be classified (2). Five recent posts in the same thread as the original post, and (3). Five recent posts by the same user in the one week window leading up to the current post.

### Deep neural network architecture

The main component is a convolutional neural network (CNN) that processes the document matrix, where each row is the corresponding word's embedding. Additional features based on posts in the thread history and user history were processed through long short-term memory (LSTM) networks and added as features as inputs to the final softmax layer of the CNN along with non n-gram features features from Traditional Machine learning features.

For the CNN component, the input is the textual content of a post represented as a sequence of words $w = (w_1, w_2, \ldots, w_n)$ each represented by their corresponding index to the vocabulary $V$. The words are mapped to word vectors via an embedding matrix $E \in \mathbb{R}^{|V| \times d}$ to produce a document matrix $D \in \mathbb{R}^{n \times d}$ where $d$ is the dimension of the word representation vectors. The central idea in CNNs is the so called *convolution* operation over the document matrix to produce a feature map representation using a *convolution filter* (CF). The goal is to learn multiple CFs that can collectively capture diverse representations of the same document. Choosing $k$ filters results in $k$ corresponding feature maps $\mathbf{v}^1, \cdots, \mathbf{v}^k$. The most distinctive feature of each feature map is selected using a max-over-time pooling operation (Collobert et al., 2011) to produce the final feature vector $\mathbf{p} \in \mathbb{R}^k$, such that $\mathbf{p} = [v^1_{max}, \cdots, v^k_{max}]$ where $v^j_{max} = max(\mathbf{v}^j_1, \cdots, \mathbf{v}^j_{n-h+1})$. Different sets of $k$ CFs are learned for different window sizes $h$ as is typically the practice. The window sizes are parameterized as a sequence $h_1, \cdots, h_H$ of $H$ unique sizes. Suppose $\mathbf{p}^{h_i}$ denotes the feature vector produced on $k$ filters with a window size of $h_i$, then the final $kH \times 1$ feature vector is

$$\mathbf{p}^* = p^{h_1} || \cdots || p^{h_H} \tag{4.1}$$

where $||$ is the vector concatenation operation. The resulting vector $\mathbf{p}^*$ provides a semantic representation of the post. Prior efforts (Kim, 2014b; Rios and Kavuluru, 2015) focus on additional details CNNs for text classification. Next, additional wide features that are concatenated to $\mathbf{p}^*$ are described.

To represent the history posts, as a preparation step prior to training, document embedding vectors over the entire corpus (even for documents without labels) are produced. This is done by first producing sparse document representations using uni/bigram counts and then applying dimensionality reduction (namely, SVD) to obtain dense vector representations that are composed using an LSTM. Suppose for the target post, the thread history consists of post embeddings $\mathbf{t_1}, \cdots, \mathbf{t_5}$ and the author's post history consists of post embeddings $\mathbf{a_1}, \cdots, \mathbf{a_5}$ as ordered from earliest to latest. We compose the post history as

$$\begin{aligned} \overrightarrow{\mathbf{t}}^i &= LSTM^{\rightarrow}(\mathbf{t_i}) \ for \ i = 1, \cdots, 5, \\ \overrightarrow{\mathbf{a}}^i &= LSTM^{\rightarrow}(\mathbf{a_i}) \ for \ i = 1, \cdots, 5, \\ \mathbf{h}^* &= \overrightarrow{\mathbf{t}}^5 \, || \, \overrightarrow{\mathbf{a}}^5 \end{aligned} \tag{4.2}$$

where $LSTM^{\rightarrow}$ represents a forward-LSTM unit composition and $\mathbf{h}^*$ is the composed vector representation of historical posts.

Besides CNN feature maps and LSTM history representations, additional features (except for $n$-gram and word embeddings) from Section 1.1 are also included, collectively represented by vector $\mathbf{g}^*$. Then the final feature vector of the full network is composed via concatenation:

$$\mathbf{f}^* = \mathbf{p}^* || \mathbf{h}^* || \mathbf{g}^* \tag{4.3}$$

The output layer $\mathbf{q} \in \mathbb{R}^m$ consists of fully-connected units that correspond to the m target labels for prediction. A softmax function is used to obtain a categorical distribution over the label space. This layer is defined as

$$q_i = \frac{e^{\mathbf{s}_i}}{\sum_{j=0}^{m} e^{\mathbf{s}_j}}, \quad \mathbf{s} = W_q \cdot \mathbf{f}^* + b_q, \tag{4.4}$$

where $W_q \in \mathbb{R}^{m \times length(\mathbf{f}^*)}$ is a parameter matrix and $b_q \in \mathbb{R}^m$ is the vector of bias terms.

### Shared task model

During the competition, we created numerous models with traditional machine learning methods and recent deep learning approaches. Our best model during the shared task was combining traditional machine learning described in the section 4.2.10 and deep learning methods described in section 4.2.10. Moreover, we found in general traditional machine learning approaches are doing better with recall and deep learning approaches perform better on precision. The combination between them made our model more robust and achieved better trade-offs between precision and recall.

### Current best model

In recent years, applying feature selection methods in healthcare datasets have been on the rise (Harb and Desuky, 2014; Sasikala et al., 2016; Jain and Singh, 2018). The challenge is how to obtain the optional set of features without increasing model complexity. The search space for $m$ features is $O(2^m)$ to explore all possible combinations. Thus exhaustive search is impractical unless $m$ is small. This problem is an NP-hard question (Guyon and Elisseeff, 2003). Therefore we used the wrapper model to make the feature selection instead of the filter model.

**Algorithm 1** Feature Selection

```
Result: Find an optimal subset SU
input  : features set: $M$, classifier model: F, variance of features: $V_m$, threshold: t
output: SU
$best\_score = 0$, $t = 0.5$
$T_n$ is min threshold for feature variance
while $t \geq T_n$ do
    $subset$ = [$m_i$ for $m_i$ in $M$ where $V_{m_i} \geq t$ ]
    $score = F(subset)$
    if $score \geq best\_score$ then
        $best\_score = score$
        SU = $subset$
    end
    $t = t/2$
end
```

The algorithm 1 shows our approach to find the optimal subset of features to train the classifier model. The *score* is based on the F-1 score with cross-validation on the training data. After finding the optimal subset, we trained our ensemble traditional

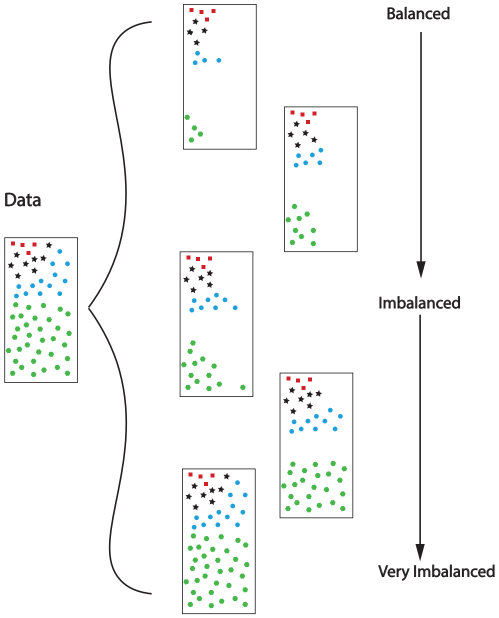


Figure 4.2: Incremental under sampling for handling imbalanced scenarios

classifier to predict on the testing data. The main takeaway from the algorithm is that instead of exhaustively searching through every possible combination in the feature space, we only search through the space via a fixed number of iterations that go through the feature standard deviation thresholds. This can be controlled by choosing an appropriate $T_n$ where smaller values lead to more steps. Thus we obviate the issues arising from handling exponential number of subsets of the feature set.

The next challenge for this task is lack of large training datasets, especially the most critical group that needs immediate help. To solve this imbalance issue with a small amount of labeled data, we proposed two undersampling approaches, which are incremental undersampling and random undersampling. Tradition undersampling sampling technique reduces the sample size of majority classes; therefore, the dataset becomes balanced. In our approach, we start with the balanced dataset, and then incrementally add more samples from majority classes. The distribution changes from balanced to slightly imbalanced until we reach the original class distribution.

Figure 4.2 provides a high-level overview of the incremental undersampling approach. The most similar approach to this is by Zhu et al. (Zhu et al., 2013) where they used random undersampling. Each time they pick a subset $SL_i$ from whole label data $D$ and train a model based on $SL_i$ and applied the model on unlabeled data $U$. Combine all the output from each model to get the final results.

Algorithm 2 represents incremental undersampling for the class prediction. The idea behind this incremental undersampling method is each subset $SL_i$ provides a different distribution from balanced to imbalanced. Since each time we shuffle the non-minority class and pick subsets from them, comparing with modifying the weight for the post approach, the undersampling method can reduce the affection of the noise data to the classifier model. Besides, the different class distributions can increase the model generalize ability on the unseen dataset.

**Algorithm 2** Incremental under sampling for high imbalance datasets

```
input : Training set $D$ where $D_{amber}, D_{crisis}, D_{green}, D_{red}$ represent subsets for each
        class respectively, $N$: number of incremental steps, Testing set $D^{test}$
output: Prediction labels
$D_{current} = D_{crisis} \cup \{|D_{crisis}|$ examples from all other classes$\}$
for $n$ from 1 to $N$ do
    $D_{current} = D_{current} \cup \{n \cdot D_l/N$ new examples from non-crisis class $l\}$
    Train stacking classifier $clf$ using $D_{current}$
    Get prediction probabilities $P_n$ for $D^{test}$
end
$Prob\_estimates(D^{test})$ = Average $N$ prediction probabilities $P_1, \ldots, P_N$
Derive labels from $Prob\_estimates(D^{test})$
```

### 4.3 Experiments

Our experiments range from traditional machine learning method to deep neural networks, from a single model to complex ensembles. The traditional machine learning model includes Support Vector Machine (SVM), Logistic Regression (LR), K-nearest neighbors k-NN).

The Table 4.2 compares traditional models; the result was computed based on one hundred times shuffling of the training dataset to compute cross validated score. The undersampling method brings the most variants; the reason is each time the selected subset was different. The contribution of each example to the model varies; therefore, we used averaged F1-score with confidence interval to measure the model performance. In this table, the top five rows involve all the features, and the bottom rows in italics are those with feature selection by removing features whose standard deviation is less than 0.0075. We believe that a lower standard deviation indicates that the feature contains less information for the model to learn. In the test phase, the stack model represents that we applied the incremental undersampling technique to that model. As shown in the results, the undersampling method can be beneficial to LR. The potential reason for it to work not as well in SVM model is this model is based on the support vectors instead of all the training data. The weak classifier trained based on the subset is too weak, and stacking them is not yielding a robust classifier. Since LR outperforms other single models in the training phase, instead of treating each model equally, we double the weight for the LR model in the stacking model.

Table 4.2: Traditional model comparison for message triaging

| | Train Set | | Test Set | |
|---|---|---|---|---|
| | Averaged F1 | 95% CI | Stack Model± 95%CI | Single Model |
| KNN | 32.61% | 32.61% ± 0.21% | 20.51% ± 0.18% | 21.00% |
| LR | 45.13% | 45.13% ± 0.20% | **47.13% ± 0.18%** | 41.57% |
| SVM-rbf | 41.41% | 41.41% ± 0.20% | 34.54% ± 0.28% | 45.94% |
| SVM-linear | 41.08% | 41.08% ± 0.19% | 34.57% ± 0.30% | 45.94% |
| Stacking | **45.68%** | 45.68% ± 0.19% | 46.81% ± 0.15% | 41.51% |
| *KNN* | 32.73% | 32.73% ± 0.19% | 20.27% ± 0.13% | 21.00% |
| *LR* | 45.17% | 45.17% ± 0.21% | **47.14% ± 0.15%** | 43.04% |
| *SVM-rbf* | 37.30% | 37.30% ± 0.24% | 24.30% ± 0.16% | 28.95% |
| *SVM-linear* | 41.38% | 41.38% ± 0.21% | 34.66% ± 0.32% | 45.94% |
| *Stacking* | **45.51%** | 45.51% ± 0.20% | 46.96% ± 0.15% | 41.51% |

During the shared task, none of the teams using unlabeled posts besides training word vectors and topic modeling. In the section, we used a semi-supervised learning approach named co-training (Blum and Mitchell, 1998) to take advantage of the unlabeled dataset. The idea is training two classifiers $C_1$ and $C_2$ on two different feature sets $v_1 \subseteq V$ and $v_2 \subseteq V$, where $V$ is the set of all features. Next, we apply $C_1$ and $C_2$ on the unlabeled dataset, and pick the most confident $p$ positive samples and $n$ negative samples from each classifier. Adding $2p$ and $2n$ examples to the labeled dataset, we repeat this for $k$ steps. Since the unlabeled dataset contains over 100,000 posts, it is inefficient to make predictions on all the posts and only pick a few of them as self-labeled samples. Therefore, we adopt the idea of batch normalization (Ioffe and Szegedy, 2015) by splitting the whole unlabeled posts to different batches with size of each equal to that of the labeled dataset. The mini-batch co-training approach can boost the speed by only predicting on a part of data. On the contrary, the original co-training used the approach by adding new examples to unlabeled set $U'$ to replace the self-labeled samples. The algorithm for Mini-Batch-Co-training is described in Algorithm 3.

```
Algorithm 3 Mini-Batch-Co-training
Given: Unlabeled samples $D^U$, subset unlabeled samples $D^{SU}$ initially labeled sample $D^L$, most positive confident examples $p$, most negative confident example $n$.
for i from 1 to 100 do
    Train Classifier1 (sentiment features) using $D^L$
    Train Classifier2 (all other features) using $D^L$
    Allow Classifier1 to label $p$ positive, $n$ negative example from $D^{SU_i}$
    Allow Classifier2 to label $p$ positive, $n$ negative example from $D^{SU_i}$
    Add $2p$+$2n$ most confident self-labeled examples to $D^L$
end
```

Our experiment shows in Table 4.3 that co-training outperformed the stacking model by 2% improvement in Flagged vs. Green task. We believe if people have anxiety, suicidal thoughts, their posts typically contain language conveying negative sentiments. On the contrary, the people who do not have mental issues, the posts are usually neutral or positive. This makes it look similar to sentiment analysis. Therefore, in terms of views for co-training, for one view we used sentiment lexicon-based features, and for the second view we used all other features. The experiment setup as following where we trained different models with and without co-training approach.

Table 4.3: Model comparisons with the co-training approach

| | Without CoTrain | | | With CoTrain | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Voting | 88.76% | 85.87% | 87.29% | 88.64% | 84.78% | 86.67% |
| RF | **91.87%** | 79.89% | 85.47% | **93.33**% | 76.09% | 83.83% |
| LR | 87.30% | **89.67%** | 88.47% | 86.84% | 89.67% | 88.24% |
| SVM-rbf | 79.90% | 84.24% | 82.01% | 80.31% | 84.24% | 82.23% |
| SVM-linear | 86.63% | 88.04% | 87.33% | 86.02% | 86.96% | 86.49% |
| Stacking | 88.24% | **89.67%** | 88.95% | 91.21% | **90.22**% | **90.71**% |
| CoTrain LR | 89.56% | 88.59% | **89.07%** | | | |

As we can see comparing co-training with each model, the co-training has the best F1-score with 89.07%. We can also combine the co-training with each model, and we found co-training can be beneficial for the stacking model with almost 2% improvement.

## 4.4 Conclusion and Future work

In this chapter, we created different models for triaging mental health messages needing immediate moderator attention in the context of an online support forum. We first described our shared task model, which combines the ensemble model of traditional ML methods with the LSTM model achieving an F1-score of 44.7%. After the shared task, we continuously improve our model through feature selection and building an ensemble model involving traditional ML methods. We obtain over 2% improvement resulting from the current best model with F1-score 47.14%. In addition, we also explore involving unlabeled data with co-training. Our preliminary experiment indicates that the co-training technique results in a 2% improvement in F1 measures with a score of 90.71% for classifying the non-green instances in the binary setup.

In the future, we will continue to explore semi-supervised learning. Our result had shown that the co-training technique performs well in the balanced dataset, but when we applied this approach to other subtasks, it has a low performance. This could be because co-training's effectiveness depends on the views chosen to train the two classifiers. For the non-green task we were able to identify sentiment features forming a useful view. For other subtasks, we will need to find other feature spaces to separate the degree of negative sentiment. For instance, the words *kill, suicide* are worse (in negative sentiment) than *unhappy, annoying* in identifying severity of mental trauma. Therefore, finding new views will be critical going forward.

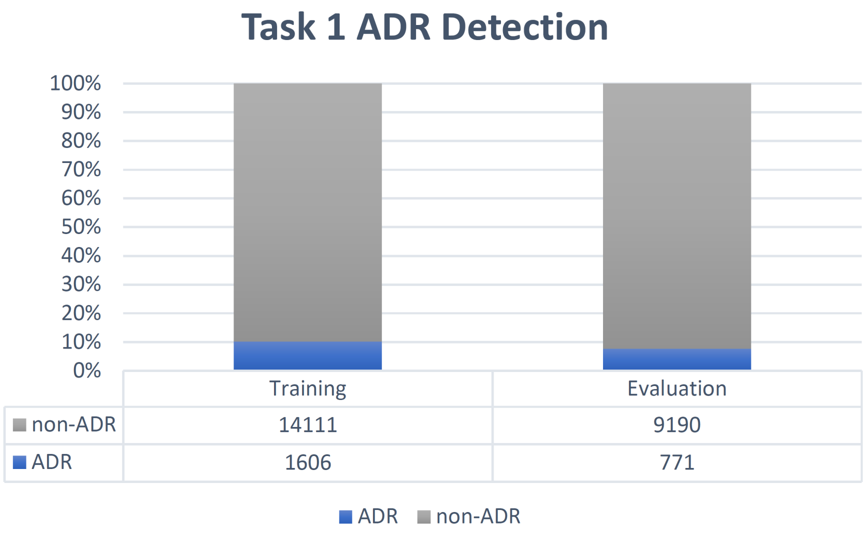
# Chapter 5 Detecting ADRs and Classifying Medication Intake Messages on Twitter

Online social networks and forums provide a new way for people to share their experiences with others. Nowadays, patients also share their symptoms and treatment progress online to help other patients dealing with similar conditions. Due to the individual differences and other factors that cannot be tested out during clinical trials, patients may experience adverse drug reactions (ADRs) even when taking FDA approved medications. ADRs lead to a financial burden of 30.1 billion dollars annually in the USA (Sultana et al., 2013). The automatic detection of ADRs is receiving significant attention from the biomedical community. While traditional approaches of reporting to the FDA (phone, online) are still crucial, given millions of patients are sharing their drug reactions on social media, automatic detection of ADRs reported on online posts may offer valuable complementary signals for pharmacovigilance. Among these posts, some ADRs are mentioned from personal experience, while other ADR mentions are observed in other people. To identify whether or not a Twitter user has consumed the medication is also an important task. So is the normalization of different ways of expressing the same event using standardized terminology. These are the three tasks in the 2nd Social Media Mining for Health Applications Shared Task at AMIA 2017. In task 1, there are 6822 training, 3555 development, and 9961 testing samples; for task 2, there are 7617 training, 2105 development, and 7513 testing samples. The first task is a binary classification task with F-score for the ADR (positive class) as the evaluation metric. Task 2 is a 3-way classification task where the evaluation metric is the micro-averaged F-score for “intake” and “possible intake” classes.

## 5.1 Materials and Methods

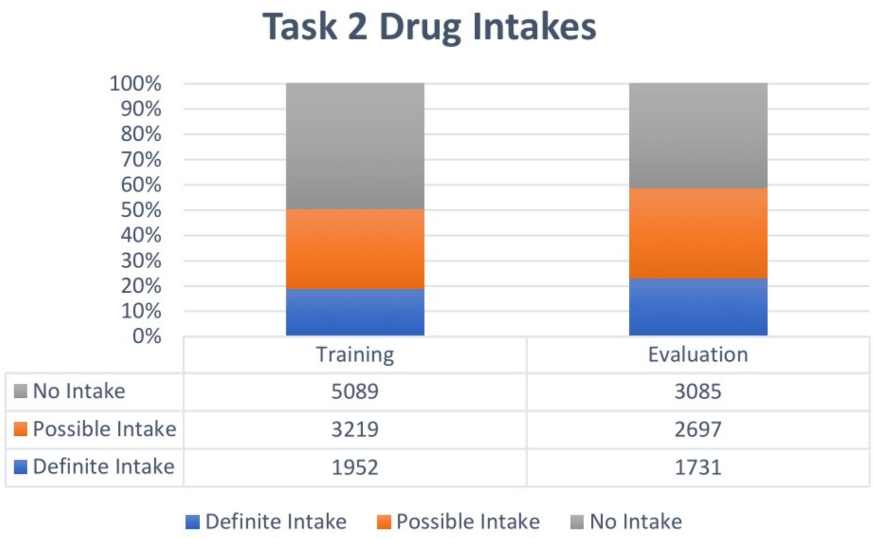
The dataset for the task 1 included 25,678 tweets annotated by the two annotators with inter-annotator agreement (IAA) of $\kappa = 0.69$ (Cohen et al., 2012). Task 2 included 17,773 annotated tweets categorized into three classes - *definite intake*, *possible intake*, and *no intake.* IAA for task 2 was $\kappa = 0.88$. Figures 5.1 and 5.2 represent the data distribution of the training and evaluation sets for tasks 1 and 2 respectively.

For tasks 1 and 2, we employ traditional methods (e.g. SVMs) and recent deep learning methods (e.g. CNNs) as well as their ensembles as detailed in Section 5.1.1 and Section 5.1.2.



| | Training | Evaluation |
|---|---|---|
| ▨ non-ADR | 14111 | 9190 |
| ■ ADR | 1606 | 771 |

Figure 5.1: Task 1 adverse drug reaction class distribution



| | Training | Evaluation |
|---|---|---|
| ▨ No Intake | 5089 | 3085 |
| ■ Possible Intake | 3219 | 2697 |
| ■ Definite Intake | 1952 | 1731 |

Figure 5.2: Task 2 drug intakes class distribution

### 5.1.1 Traditional linear models (TLMs)

The traditional machine learning requires feature engineering. In task 1 to identify the adverse drug reaction mentions, we found that domain knowledge about drugs and reactions plays an important role. For task 2 on medication intakes, we found that sentiment associations, PMI scores for the ADR lexicon, and handcrafted lexical pairs of drug mentions contribute to the traditional models. While the deep learning approach in this study showed promising results without manually feature engineering, understanding the potential of traditional features can further help build a complex model.

For tasks 1 and 2, we used both TLMs, specifically linear SVMs and logistic regression, and deep nets specifically CNNs. In the first task, we averaged the probability estimates from each TLM classifier with and without CNN ensemble. The features used in TLMs are itemized below.

- Uni/bi grams: Counts and real values of uni/bi-gram features in the tweet via *countvectorizer* and *tfidfvectorizer*, respectively, from Scikit Learn machine learning package (Pedregosa et al., 2011b).

- Sentiment lexicons (11 features): Four count features from the post based on the positive and negative counts of unigrams and bigrams using the NRC Canada emoticon lexicon[1], four additional sentiment score aggregation features corresponding to the count features, and overall sentiment score of unigrams/bigrams/uni+bi grams.

- Word embeddings (400 features): The average 400 dimensional vector of all the word vectors of a post where the word vectors ($\in \mathbb{R}^{400}$) are obtained from a pre-trained Twitter word2vec model (Godin et al., 2015).

- ADR lexicon (Sarker and Gonzalez, 2015) (7423 features): One Boolean feature per concept indicating whether the concept ID is mentioned in the tweet; rest are count features identifying the number of drugs from a particular source (SIDER, CHV, COSTART and DIEGO_lab) and the number of times different drugs are mentioned in the tweet.

- Negation words (2 features): The first is a count of certain negation words (not, don't, can't, won't, doesn't, never) and second feature is the proportion of negation words to the total number of words in the post.

- PMI score (1 feature): The sum of words' pointwise mutual information (PMI) (Bouma, 2009) scores as a real-valued feature based on the training examples and their class membership.

- For task 2 only – handcrafted lexical pairs of drug mentions preceded by pronouns (6 features): The count of first, second, and third personal pronouns with

[1] http://saifmohammad.com/WebPages/AccessResource.htm

and without negation followed, with potentially other intermediate words, by a drug mention[2]. (e.g., “I did a line of cocaine”)

The TLM ensemble was based on two logistic regression models with different C parameters and one SVM model.

### 5.1.2 Deep learning approach

In the deep learning approach we used CNN models where each tweet is passed to the model as a sequence of words vectors, $[\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n]$, where $n$ is the number of words in the tweet. We begin by concatenating each window spanning $k$ words, $\mathbf{x}_{j-k+1}|| \dots ||\mathbf{x}_j$, into a local context vector $\mathbf{c}_j \in \mathbb{R}^{kd_{emb}}$ where $d_{emb}$ is the dimensionality of the word vectors. Intuitively, CNNs extract informative n-grams from text and n-grams are extracted with the use of *convolutional filters* (CFs). We define the CFs as $\mathbf{W}_k \in \mathbb{R}^{q \times kd_{emb}}$, where $q$ is the number of *feature maps* generated using filter width $k$. Next, using a non-linear function $f$, we convolve over each context vector,

$$\hat{\mathbf{c}}_j = f(\mathbf{W}\mathbf{c}_j + \mathbf{b}),$$

where $\mathbf{b} \in \mathbb{R}^q$. Given the convolved context vectors $[\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \dots, \hat{\mathbf{c}}_{n-k+1}]$, we map them into a fixed size vector using max-over-time pooling

$$\mathbf{m}_k = [\hat{c}^1_{max}, \hat{c}^2_{max}, \dots, \hat{c}^q_{max}],$$

where $\hat{c}^j_{max}$ represents the max value across the $j$-th feature map such that $\hat{c}^j_{max} = max(\hat{c}^j_1, \hat{c}^j_2, \dots, \hat{c}^j_{n-k+1})$. To improve our model, we use convolutional filters of different widths. With filters spanning a different number of words $(k_1, k_2, \dots, k_p)$, we generate multiple sentence representations $\mathbf{m}_{k_1}, \mathbf{m}_{k_2}, \dots, \mathbf{m}_{k_p}$, where $p$ is the total number of filter widths we use.

When datasets contain many noisy features, it can be beneficial to use a simpler model. Simpler models are less prone to overfitting to the noise in the dataset. We augment the features generated from the CNN with a simple average of all the word vectors in a given tweet.

$$\mathbf{m}_{bow} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_i.$$

Now we have $p+1$ feature representations of the final tweet $[\mathbf{m}_{k_1}, \mathbf{m}_{k_2}, \dots, \mathbf{m}_{k_p}, \mathbf{m}_{bow}]$. Prior work using CNNs (Rios and Kavuluru, 2015) simply concatenated each $\mathbf{m}_j$ to

[2] https://www.drugs.com/drug_information.html

pass to the final fully-connected softmax layer. Rather than concatenating each $\mathbf{m}_{k_j}$, we use self-attention (Lin et al., 2017) (with multiple attention scores per feature representation) to allow the model to dynamically weight the feature representations. Specifically, we define the $j$-th attention of the $i$-th feature representation as

$$\alpha_{j,i} = \frac{exp(e_{j,i})}{\sum_{k=1}^{p+1} exp(e_{j,k})}, \text{ where } e_{j,i} = \boldsymbol{v}_j^T \cdot \tanh(\mathbf{W}_a \mathbf{m}_i),$$

such that $\mathbf{W}_a \in \mathbb{R}^{t \times q}$, $\mathbf{v}_j \in \mathbb{R}^t$, and $\alpha_{j,i} \in [0, 1]$. Intuitively, $\alpha_{j,i}$ represents the importance we give the feature representation $\mathbf{m}_i$. Next, we merge all feature representations into a matrix $\mathbf{M} \in \mathbb{R}^{(p+1)\times q}$. Likewise, given $s$ total attentions, each attention weight is combined to form an attention matrix $\mathbf{A} \in \mathbb{R}^{s\times(p+1)}$. Here, the $j$-th row of A represents the importance of each $\mathbf{m}_i$ with respect to the attention weights $\mathbf{v}_j$. Finally, we represent each tweet as the weighted sum of the feature representations according to each of the attention weight vectors as

$$\mathbf{h} = vec(\mathbf{AM}),$$

where $vec$ represents the matrix vectorization operation and $\mathbf{h} \in \mathbb{R}^{sq}$. Lastly, $\mathbf{h}$ is passed to a final fully-connected softmax layer. Figure 5.3 represents the attention based CNN which we used in both tasks.

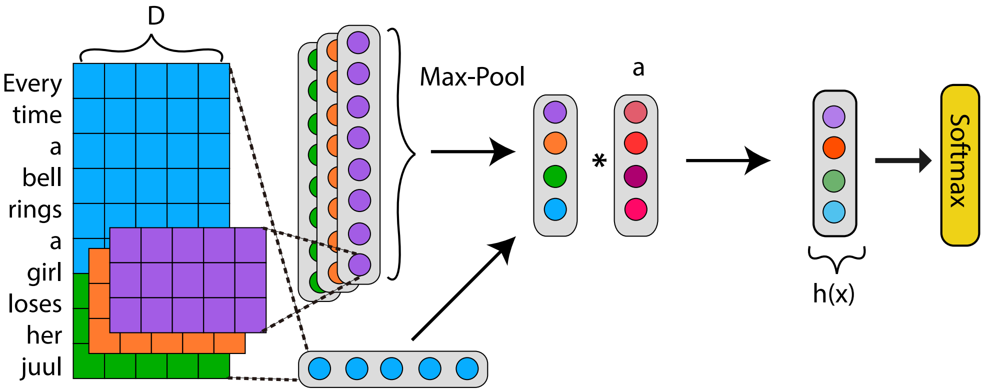


Figure 5.3: Architecture of attention based CNN model

## 5.2 Experiments

In this section, we describe the experiments we did to find the ideal model during the shared task. We used the organizer provided training and development datasets and the class distribution listed in Tables 5.1 and 5.2 respectively.

Table 5.1: Task 1 class distribution

| | Training | Development |
|---|---:|---:|
| ADR mention | 730 | 242 |
| Non-ADR mention | 6092 | 3313 |

Table 5.2: Task 2 class distribution

| | Training | Development |
|---|---:|---:|
| Definite Intake | 1462 | 397 |
| Possible Intake | 2383 | 667 |
| No Intake | 3772 | 1041 |

We trained our models based on the training set and evaluated them on the development dataset to find the candidate models for the submission. The parameters used for the CNN included convolution filters spanning 3,4, and 5 words, and 400 filters of each span length. We used a dropout rate of 0.5 and batch size 50. The dimensionality for word vectors used was 400. The optimizer used in this CNN model is Adam, and the loss function is cross-entropy loss. The C parameters for logistic regression were 0.3 and 0.3 with class weight set as *balanced*.

## 5.3 Results and Discussion

### 5.3.1 Task 1 results

Table 5.3 shows our official scores on task 1. The CNN with Attention (CNN-Att) model is observed to outperform the other models on F1-score measurement. On the training and development data set, we found logistic regression to perform better than SVM; TLM ensemble (model averaging) of two LR models with one SVM model and the base CNN have about the same performance while the averaged model of TLM-ensemble with CNN was the best performer. CNN-Att itself was slightly better than TLM-ensemble and CNN when considered separately, but CNN-Att was worse off when TLM-ensemble and CNN were combined using model averaging. Our ensemble model has the top precision score among all teams and shows the potential of the ensemble approaches. However, our recall is significantly less than the best performer where it was over 48%. We further investigate the discrepancies between train and test set performances of various models. Our initial assessment is that the attention model is able to weight what different CFs are capturing from each tweet more effectively. To verify this, we examine the "popular" $n$-grams of filters (based on values of $\hat{\mathbf{c}}_j$ in Section 5.1.2) that are consistently being weighted above others by the attention mechanism. In turn, these $n$-grams can be used as additional features in the regular CNN or TML models to potentially improve their performances. We need to experiment with these additional approaches to improve our recall without significant compromises in precision.

Table 5.3: Task 1: performance on the test set

| | ADR Precision | ADR Recall | ADR F-score |
|---|---:|---:|---:|
| Baseline 1: Naive Bayes | 0.774 | 0.098 | 0.174 |
| Baseline 2: SVMs with RBF kernel | 0.501 | 0.215 | 0.219 |
| Baseline 3: Random Forest | 0.429 | 0.066 | 0.115 |
| TLM-ensemble | 0.459 | 0.237 | 0.313 |
| CNN+TLM-ensemble | **0.567** | 0.259 | 0.356 |
| CNN-Att | 0.498 | 0.337 | **0.402** |

### 5.3.2 Task 2 results

Table 5.4 shows our official scores on task 2. The CNN-Att model still has the best performance over the other two models we submitted. On the training and development data sets, we found deep nets significantly outperformed the TLMs (and their ensembles). Therefore, in task 2, we focused on incorporating deep nets in all ensembles.

Table 5.4: Task 2: performance on the test set

| | Prec. for classes 1 & 2 | Recall for classes 1 & 2 | F-score for classes 1 & 2 |
|---|---:|---:|---:|
| Baseline 1: Naive Bayes | 0.359 | 0.503 | 0.419 |
| Baseline 2: SVMs | 0.652 | 0.436 | 0.523 |
| Baseline 3: Random Forest | 0.628 | 0.487 | 0.549 |
| CNN+TLM-ensemble | 0.688 | 0.607 | 0.645 |
| CNNs | 0.705 | 0.666 | 0.685 |
| CNN-Att | 0.701 | **0.677** | **0.689** |

Since each task allowed submission of only three models, the ones we submitted may not have been the best compared with additional models we built as part of our participation in the challenge. Therefore, we did some new experiments using unsubmitted models and found newer,better-performing ensembles on tasks 1 and 2 as shown in Table 5.5. The results from our CNN and CNN-Att models are from averaging ensembles of ten models, each using both stratified and random 80%–20% splits. We did this because the test dataset distribution may vary from training data, and we believed stratified split might cause overfitting. Therefore, we felt that with the help of random splits, it might help our model's ability to handle test data with a different distribution. In this particular case, the test data distribution is similar to that of the training data, so when we only consider the use of stratified splitting to tune the parameters, our results improved. In task 1, we found the ensemble model with CNN-Att and CNN with logistic regression had almost a 1% improvement in the F1 score. In task 2, we found CNN-Att trained with stratified splitting has an F1 score that is 0.4% higher than that of our submitted model, and this score is equal to the winning team's performance.

Table 5.5: Experiments on unsubmitted models

| | Precision | Recall | F-score |
|---|---|---|---|
| CNN_Att+CNN+LR (task1) | 0.483 | 0.358 | 0.411 |
| CNN_Att (averaging_stratified_only) (task2) | 0.701 | 0.686 | 0.693 |

### 5.3.3 Post-task result

During the shared task, due to to Twitter's restrictions, the organizers provided only the tweetIDs for participants to download the tweets. Since some users already deleted their tweets or unregistered their accounts, not all labeled tweets were available to use. After the competition, all the participate teams were invited to write a joint paper to describe the models and findings. The organizer shared the full dataset to all the participant teams in the context. We retrained our model, and our current best model performance got further improvements; the results are shown in Table 5.6. In task1, our best model is an ensemble model of CNN-Att with logistic regression,

Table 5.6: Model performance on full dataset

| Model | Subtask (evaluation metric) | F-score | Performance change |
|---|---|---|---|
| CNN_Att+LR | 1 (ADR F1-score) | 0.459 | +0.057 |
| CNN_Att | 2 (micro F1-score for classes 1 & 2) | 0.694 | +0.005 |

given the traditional model was more precise and the deep learning model had better recall. On the full dataset we gained another 5.7% improvement. In task 2, we applied a small modification during the pre-processing task by removing all non-ASCII characters from tweets and applied CNN-Att alone to get the best performance among all teams. The improvement in task 2 is only 0.5%, which indicates that the volume of data cannot help very much in this case. For further studies, we need to think about the architecture of the deep learning model. For instance, a wide and deep model in which we can add some useful handcrafted features to train the model maybe better.

## 5.4 Discussion

In this section, we start with some error analysis to study the weaknesses of the current model and provide a guide for the future research directions. We end with some model comparisons to understand the performance differences between the traditional machine learning approach and deep neural networks on the different tasks.

### 5.4.1 Error analysis

In both tasks, false negatives were the use of the infrequent, creative expressions. For instance, *i have metformin tummy today:-(* and the rear ADRs are usually misclassified. False positive occurs when classifiers cannot correctly extract the relation between drug and its adverse effects such as the symptom *headache* but instead link to its positive effects such as *hair loss reversal.* Besides, the IAA score for task 1 is 0.69, which shows this is a more difficult task since human beings are also having a difficult time understanding the meaning. This is a common challenge in the Twitter data analysis; due to the short informal language, it is not uncommon to misunderstand. In task 2, there are implicit mentions about medication usage and sometimes explicit mentions about it; at times, however, it is unclear whether the drug usage is carried out by the tweet author. This is a known challenge in analyzing tweets as Twitters users may express opinions across multiple posts. Therefore, more sophisticated grouping of tweets can future improve the overall performance.

Another type of misclassification occurs when the concepts are closely related, such as *Insomnia* and *Somnolence* or when handling antonymous concepts such as *Insomnia* and *Hypersomnia.* In the phrase *"sleep for X hours,"* only the actual number of hours people specify can differentiate the two concepts regardless of the rest of the message. This requires additional domain knowledge and is challenging for machine learning models to learn without such explicit input. With limited training data, rarely occurring concepts are difficult to correctly classify. For example, *Night sweats* were misclassified as *Hyperhidrosis*, and there two instances of it in the training data without explicit mentions of the keyword *night* (e.g., *"waking up in a pool of your own sweat"*). Overall, error analysis provides new guidelines for future model design and improvements.

### 5.4.2 Model comparison

In both tasks, the deep learning approaches outperform traditional machine learning methods. One reason is because SVM and logistic regression rely on feature engineering. In this task, explicit domain knowledge about medical concepts and the corresponding symptoms would be helpful to improve model performance. Our current linear model features included several lexicon lists, which show the effectiveness of domain-specific knowledge. On the contrary, the deep learning approach allows the model to capture the features more automatically. This gives a lot of freedom for the model to find the features that are useful to it, even if the features are not as interpretable to humans. With the attention mechanism applied, it increased the model's ability to focus on the most important features given not all features have an equal contribution to the classifier. In addition, given the performances of the models during the shared task and post-task, we found that the deep learning's effectiveness relies to a large extent on the size of the training dataset where a large dataset lets the model identify more latent features and helps tune the parameters more effectively.

## 5.5 Conclusion

In this chapter, we applied NLP methods for two tasks. The first task is detecting the adverse drug reaction mentions from a short text. This task presents a common challenge in the healthcare field, where is data imbalanced. The key difference between this chapter with chapter 4 is the size of training data and the length of each post. As shown in section 5.3 and discussion, deep learning outperforms traditional machine learning models on this task. It demonstrates the ability of deep models in handling large datasets. Although our model gets the best performance, it still shows this is a very challenge task. With the low precision and recall, we need to involve more domain knowledge and create a more complex model to better perform on this task.

The second task in this chapter is to identify if the author of the message is using a medication. The data distribution is much more balanced compared with that of task 1. The challenge in this task arises in implicit mentions of the usage of the medication and lack of clarity in who used the drug (the tweet author or someone else not mentioned in the tweet).

## Chapter 6 Identifying Juul Users from Tweets mentioning Juul

Electronic cigarettes (E-cigs) are popular among youth and are being marketed as an alternative to tobacco-based cigarettes to quit smoking. They are often marketed to be risk-free for new users. According to Nielsen data (Herzog and Kanada, 2018), it is estimated that the market of E-cigs has grown to $5.5 billion in 2018 from $1.7 billion as reported by CNBC in 2013 (Mangan, 2015). Juul, an e-cig brand, represented 75% of the market in 2018. As such it is very popular and is often discussed on different social network platforms.

A Juul device looks like a flash drive—it even connects to a USB port for charging, which is convenient for people to carry and use. Juul is marketed to young users by having a strong social media presence where users share their "juuling" experiences on Twitter, Instagram, and other social media. Moreover, with the abundance of different flavors provided, Juul attracts adolescent users, even if they never smoked before. A Juul pod contains the nicotine equivalent to a pack of cigarettes (Initiative, 2018a). Furthermore, nicotine adversely affects cognitive development at an early age (Goriounova and Mansvelder, 2012). In a recent report by the Truth Initiative (Initiative, 2018b), they show teenagers are 16 times more likely to use Juul than older age groups. Given the wide-spread use of social media by young adults and teenagers, marketing efforts using social media directly impacts the use of this demographic. In this chapter, we introduce a novel neural network-based method to find juulers (i.e., people who use the Juul e-cig device). Our techniques can be applied to many downstream tasks including, but not limited to, detecting adolescents at-risk of becoming juulers and used as part of epidemiological studies of e-cig use and dependence as explicitly related to Juul. Concerning epidemiological studies, Juul provides a unique case study of the impact of substantial social media marketing strategies on young people.

We are developing methods for a public health surveillance task. Social media has been used for a multitude of public health surveillance tasks. For example, Kim et al.(Kim et al., 2017) proposed a Gradient Boosting Regression Trees (GBRT) method to classify e-cig users into one of five categories: individual, vaper enthusiast, informed agency, marketer, and spammer. This model uses tweet metadata and tweeting behavior, derived from a user's recent 200 posts to classify the categories. In Kim's report, their model's performance on vaper enthusiast is 47.1% F1-score, 40% recall, and 57.1% precision. Their study is the closest to our study. Compared to Kim et al., our model differs in three substantial ways. First, we develop classifiers for tweet-level prediction, rather than using the last 200 posts for a specific user. Requiring a model only to generate predictions for users with 200 posts is not practical, especially when we are interested in users who actually use the Juul device. Tweet-based methods struggle to predict user-based factors when they only have access to a few tweets. Moreover, many users on twitter tweet infrequently. In a dataset of over 700 million users from Chamberlain et al. (Chamberlain et al., 2017), the median number of tweets per user was only four. Therefore, developing accurate tweet-level methods is important for public health surveillance applications. Second, we combine deep learning and traditional machine learning to reach a new state-of-the-art result to identify the Juul user, while Kim et al. ignore neural network-based methods. Third, Kim et al. studied E-cigs in general; however, in this work, we focus on Juul. Juul holds 75% of the E-cig market, and is popular among teenagers, justifying a study focused on Juul.

Neural Networks have produced state-of-the-art results across a wide range of text classification tasks. Han et al., 2017 used attention-based convolutional neural networks (CNNs) to identify the ADR mentions and medicine intakes in tweets, Kavuluru, Rios, and Tran (Kavuluru et al., 2017) applied word and character-level recurrent neural networks to extract drug-drug interactions, Tung et al.(Tran and Kavuluru, 2017) applied recurrent neural networks with hierarchical attention (Re-HAN) for predicting mental conditions based on psychiatric notes. Numerous works applied neural networks to solve questions in the biomedical domain (Chen et al., 2016; Li et al., 2015; Quan et al., 2016).

In this chapter, we present a novel method that combines neural networks with handcrafted features to predict whether the author of a tweet is a juuler or not.

Specifically, we make the following research contributions:

- We introduce a new publicly available dataset [1] of 1.3 million tweets about Juul where 3000 tweets have been manually annotated with each user's juuler status.
- We present a novel method for identifying juulers that combines neural networks (recurrent and convolutional) with feature engineering.
- We perform a fine-grained error analysis that explores how personal pronouns affect the classification and provide a possible solution.

[1] https://github.com/sifei/Identify-the-Juuler

## 6.1 Materials and Methods

The juuler classification model is based on Wide&Deep learning (Cheng et al., 2016) configurations. We also employed Long-Short Term Memory (LSTM) networks to help model take into account syntactic information. In this section, we discuss the model we used to classify Juul users.

### Datasets

We used twitter streaming API to collect tweets that mention different variants of Juul (juul, juuling, juuled) between 10/19/2017 and 08/17/2018. We randomly select 3000 posts from over 1.3 million tweets and asked three annotators to individually identify whether the author of the tweet is a juuler or not. Therefore, a 3000 labeled dataset created for this study, and we used majority voting to solve the annotator's disagreement - the kappa inter-rater agreement score 71.37%, which means substantial agreement. There are 1473 out of 3000 tweets created by the juuler, meaning almost half of the Twitter users who tweet about Juul are actual Juul users as per this dataset.

### 6.1.1 Method overview

The goal of the model is to build a binary classifier that output juulers create the probability estimates of given tweets. This model's idea is adopted from Cheng et al. (Cheng et al., 2016) and contains two parts: wide and deep features. The wide model uses more straightforward rules which represent how humans identify the differences between a user and a non-user. The deep model provides automatically learns discriminative features to help identify Juul users.

### 6.1.2 Wide model

The wide model used the feature engineering approach with hand-crafted features to create a generalized linear model of the form $y = \mathbf{w}^T\mathbf{x}+b$ where $y$ is the prediction and $\mathbf{x} = [x_1, x_2, ..., x_n]$ is a vector of $n$ features, $\mathbf{w} = [w_1, w_2, ..., w_n]$ is the feature weight vector learned via training and $b$ is the bias. The handcrafted features included the count of first, second, and third personal pronouns, negation words, and verbs. Besides the handcrafted features, we used spaCy (Honnibal and Johnson, 2015) for part-of-speech tagging and dependency parsing to gain syntactic features of the tweet. This part of the model helps memorize interesting tokens and syntactic features.

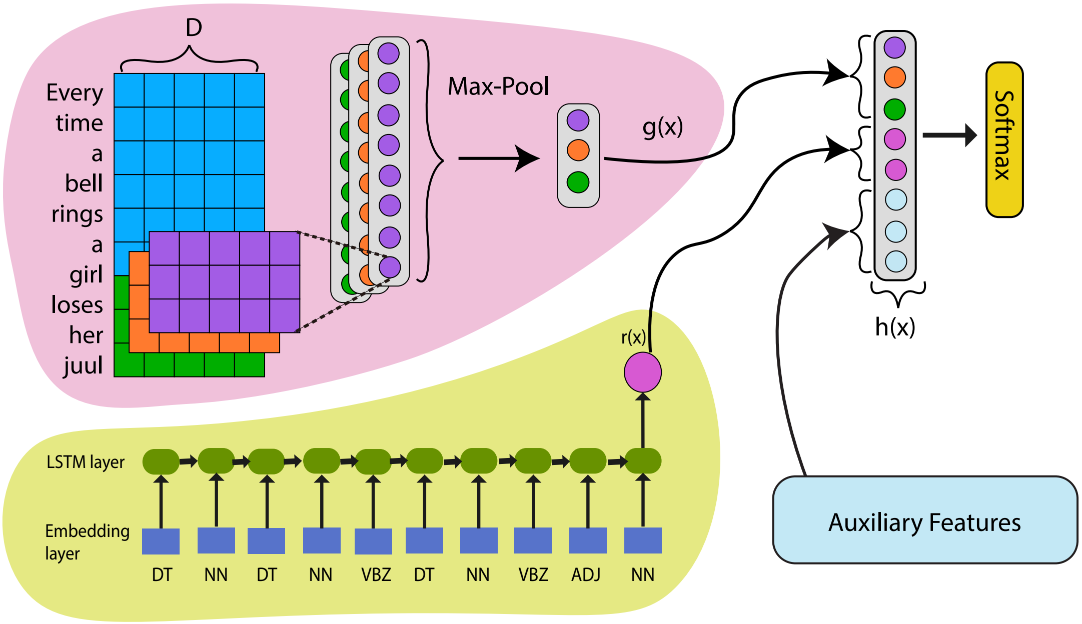


Figure 6.1: Wide and deep model

### 6.1.3 Deep model

The deep model contains two parts of the model: 1) CNNs to extract signal from the tweet text. 2) LSTMs to capture the part-of-speech relationship to the corresponding word in the tweet. The idea behind the deep model is to let the computer gain the generalize ability on the unseen words, which the model can match the close examples to identify the new post.

#### Convolutional neural networks

CNNs use convolution filters to capture the representation of contiguous terms based on the filter size. A convolution operation is applying a convolution filter on a segment of text to output one single real number represents that segment of text. As shown in the top part of the figure 6.1, the input layer takes a tweet and represents it as a matrix $\mathbf{D}_i \in \mathbb{R}^{\mathbf{x}_i \times d}$, where $\mathbf{x}_i$ is the $i$-th word in the tweet and $d$ is the size of word vector. The next step is using convolution filters transform document matrix $\mathbf{D}$ into a vector. The convolution filter size corresponds to $n$ in the n-gram features. Therefore multiple convolution filters produce various vector representations of the tweet. To form a fixed size vector representation, we use max-over-time pooling across each vector to get a single real value. Given the document matrix $\mathbf{D}_i$ the CNN produces a fixed size feature representation

$$g(\mathbf{D}_i) = CNN(\mathbf{D}_i)$$

where $g(\mathbf{D}_i) \in \mathbb{R}^{f \cdot s}$, $s$ is the number of convolution filter sizes, and $f$ is the number of filters per size.

### LSTMs

The general idea of using the part-of-speech tags via the LSTM model comes from the notion of co-training (Blum and Mitchell, 1998) by training a model on different views (feature sets) to arrive at classifiers that offer different complementary predictive power. Therefore, we used LSTM in our full model.

LSTMs are extensions of the basic RNN setup and were designed to obviate the vanishing gradient issue that plagued the regular RNNs. LSTMs use Eq 6.1 to decide how much information needs to be remembered from the previous output.

$$f_t = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_f) \tag{6.1}$$

where $f_t$ is forget gate, the hidden vector $\mathbf{h}_{t-1} \in \mathbb{R}^{a \times b}$ where $a$ is hidden dimension and $b$ is 1, the input vector $\mathbf{x}_t \in \mathbb{R}^{c \times b}$ where $c$ is embedding dimension and $\mathbf{W}_f \in \mathbb{R}^{d \times e}$ is the parameter matrix for forget gate where $d$ is batch size and $e$ is the sum of hidden and input dimensions, [] is a concatenation operation between hidden vector and input vector, and the bias of forget gate $\mathbf{b}_f \in \mathbb{R}^{d \times b}$.

In Eq 6.2 shown, the $i_t$ is input gate, similar to forget gate, this gate control how many information need to remember from the input vector where $\mathbf{W}_i \in \mathbb{R}^{d \times e}$ and $\mathbf{W}_c \in \mathbb{R}^{d \times e}$. Then we applied tanh on input vector to get $\tilde{C}_t$ which represent the current state.

$$\begin{gathered} i_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \\ \tilde{C}_t = tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C) \end{gathered} \tag{6.2}$$

Now, we let the model learn the information based on the output from forget gate and input gate. If $f_t$ is close to zero, most information from previous stages ought to be forgotten as shown in Eq 6.3.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{6.3}$$

The last step is to create output state $o_t$ with $\mathbf{W}_o \in \mathbb{R}^{d \times e}$, and get new hidden layer vector $h_t \in \mathbb{R}^{a \times b}$

$$o_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o)$$
$$h_t = o_t * tanh(C_t)$$

As shown in the middle part of Figure 6.1, each part-of-speech tag is embedded into a vector and the vectors corresponding to tags for a given tweet are processed from left to right. The LSTM layer receives two inputs: previous state output and current part-of-speech tag vector representation. At the end of the last part-of-speech tag, the output $r(x)$ is the vector representation of the whole sequence of part-of-speech tags for the tweet. For more details of LSTM please refer to Sundermeyer et al. (Sundermeyer et al., 2012).

### 6.1.4 Wide and deep model

The Wide and Deep model combines all three models, given the wide model provides prior knowledge (memorizing), but it has weak generalization ability. On the contrary, the deep model has the superior generalizabilty, but sometimes may generalize too much, which brings adverse effects to the classification. Training wide and deep models together, the errors are backpropagated to both models to learn the parameters. Figure 6.1 shows the overall architecture of the wide & deep model. The model's prediction defined as:

$$P(Y = 1|\mathbf{x}) = \sigma(\mathbf{w}_{wide}^T \mathbf{a}(x) + \mathbf{w}_{CNN}^T \mathbf{g}(x) + \mathbf{w}_{LSTM}^T \mathbf{r}(x) + \mathbf{b})$$

where $Y$ is the binary class label, $\sigma()$ is the sigmoid function, $\mathbf{w}_{wide}$ is the vector of all wide model weights, $\mathbf{w}_{CNN}$ is the vector of CNN model weights, $\mathbf{w}_{LSTM}$ is the vector of LSTM model weights, $\mathbf{a}(x)$ is the vector of wide features, $\mathbf{g}(x)$ is the vector of CNN features for tweet content, $\mathbf{r}(x)$ is the vector of LSTM feature for part-of-speech tags, and b is the bias term.

### 6.1.5 Training

To train our model, we use the cross-entropy loss define as:

$$L(p, y) = -(y \log(p) + (1 - y) \log(1 - p)),$$

where $y$ is a binary indicator (0 or 1) 1 if correct classification and $p$ is predicted probability. In this framework, we trained all outputs jointly. The goal of this project is to find out the real juulers, which means precision is more important than recall so as to not overestimate the prevalence of different findings in downstream analyses. To this end, instead of 0.5 sigmoid score as a threshold to assign the label, we choose a lower threshold at 0.4 to trade-off recall for better precision.

## 6.2 Results

### 6.2.1 Evaluation and baselines

For evaluation proposes, we trained four traditional machine learning approaches using different combinations of feature sets, a deep model, and a wide&deep model. We briefly describe the models below:

- Support Vector Machine (SVM) – This model is used across different feature sets to train a linear SVM. The model is trained using Scikit-Learn's SVC method (Pedregosa et al., 2011c). We random-search over the regularization parameter C and kernel coefficient parameter gamma, and the class weight option ("Balanced" in scikit-learn) options.

- Logistic Regression (LR) – Uses the same features as the SVM model. Random search over regularization parameter is also used here.

- Random Forest (RF) – Similar to the previous two models, this random forest model uses the same features, and we random-search over the maximum depth *max_ depth* of the tree in terms of hyperparameter selection.

- Stacking – This model is trained using Mlxtend's stacking classifier (Raschka, 2018). It uses the probability output from each model listed above as input features to a meta-classifier, an LR model.

- Deep – Combination of both the CNN and LSTM models.

- Wide&Deep – The method proposed in this chapter.

The parameters used for the model defined as convolution filter span 2, 3, 4, 5 words, and each filter size has 100 filters. We used a dropout rate at 0.3, and the batch size set as 32. The word dimension is 400, and the part-of-speech dimension is 32; the hidden dimension is 400. The optimizer used in this model is Adam, and the loss function is cross-entropy loss. The traditional machine learning parameters are shown in Table 6.1.

Table 6.1: Parameters for traditional machine learning models

| Model | Count vectorizer | Count vectorizer + hand features | TfIdf vectorizer | TfIdf vectorizer + hand features |
|---|---|---|---|---|
| LR | C=1.11329 | C=0.411315 | C=2.3387 | C=6.2095 |
| SVM | C=61.3054, gamma=0.001 | C=8.908, gamma=0.01 | C=39.347, gamma=0.01 | C=38.38, gamma=0.01 |
| RF | max depth=65 | max depth=106 | max depth=126 | max depth=31 |

### 6.2.2 Experiments

We test various machine learning methods, SVMs, LR, Random Forest, and deep learning approaches. The table shows the f1-score, precision, and recall with the corresponding confidence intervals. In our experiments, all the data is split into 70% for training, 10% for development, 20% for testing. We used 100 times stratified sampling for the traditional machine learning models. Training a deep model is more costly. Instead of training 100 models, we trained 20 models and randomly picked ten models averaging the probabilities of each model, and repeat 100 times to get the statistical results. We compared traditional machine learning approaches with different feature sets and deep models. The results are shown in Table 6.2.

Table 6.2: Model comparison for predicting Juul use through tweets

| Features | Model | F1 | 95% CI | Precision | 95% CI | Recall | 95% CI |
|---|---|---|---|---|---|---|---|
| Count | LR | 75.63% | 75.29-75.97% | 76.87% | 76.50-77.24% | 74.49% | 73.96-75.02% |
| | SVM | 75.48% | 75.14-75.83% | 77.57% | 77.19-77.95% | 73.57% | 73.03-74.12% |
| | RF | 72.15% | 71.74-72.57% | 76.62% | 76.19-77.06% | 68.28% | 67.62-68.94% |
| | Stacking | 75.98% | 75.60-76.36% | 77.99% | 77.62-78.36% | 74.13% | 73.56-74.69% |
| Count_Hand | LR | 76.72% | 76.36-77.09% | 77.89% | 77.49-78.29% | 75.63% | 75.13-76.14% |
| | SVM | 76.17% | 75.82-76.53% | 77.22% | 76.84-77.61% | 75.21% | 74.69-75.73% |
| | RF | 72.88% | 72.48-73.29% | 76.70% | 76.31-77.09% | 69.51% | 68.88-70.14% |
| | Stacking | 76.76% | 76.40-77.12% | 78.29% | 77.89-78.70% | 75.33% | 74.84-75.82% |
| TF-IDF | LR | 77.05% | 76.73-77.36% | 78.26% | 77.87-78.65% | 76.10% | 75.67-76.53% |
| | SVM | 77.08% | 76.77-77.40% | 78.26% | 77.87-78.65% | 75.99% | 75.54-76.45% |
| | RF | 72.11% | 71.78-72.45% | 71.95% | 71.48-72.42% | 72.42% | 71.78-73.05% |
| | Stacking | 77.14% | 76.83-77.46% | 77.15% | 76.74-77.56% | 77.19% | 76.76-77.63% |
| TF-IDF_hand | LR | 77.29% | 76.96-77.62% | 77.52% | 77.13-77.90% | 77.12% | 76.63-77.61% |
| | SVM | 77.32% | 76.99-77.64% | 77.68% | 77.28-78.08% | 77.00% | 76.54-77.47% |
| | RF | 71.82% | 71.45-72.20% | 72.74% | 72.20-73.28% | 71.06% | 70.43-71.69% |
| | Stacking | *77.48%* | 77.16-77.80% | 77.55% | 77.19-77.92% | 77.45% | 76.98-77.93% |
| Neural | Deep | *77.48%* | 77.33-77.63% | **81.59%** | 81.45-81.72% | 73.78% | 73.52-74.04% |
| | **Wide&Deep** | **79.06%** | 78.95-79.17% | 80.61% | 80.52-80.69% | **77.58%** | 77.39-77.77% |

We also did an ablation experiment to understand how each constituent model contributes to the full model. Table 6.3 shows the sub-model performance by removing each part as described in the methods section 6.1. We can see that the best deep model is CNN&Wide model with F1-score 77.85%. The wide features contributed 0.51% to enhance the CNN model alone performance. On the contrary, the wide features did not help the LSTM model with 0.36% worse off LSTM alone. All these models have a higher precision score than the recall score, which is expected in our study. To further examine the effect of the LSTM model, we trained a Bi-LSTM on tweet text only and using tweet text as inputs to both CNN and LSTM model. The F1-score is 75.90% and 76.56%, precision score is 75.71% and 77.02% , and recall score is 76.11% and 76.12% respectively. In the end, we used tweets embedding and part-of-speech tag embedding as inputs to train a CNN model; the model F1-score is 78.95%, precision and recall scores are 78.82% and 79.07% respectively. We discuss more ablation experiments in the discussion section 6.3.

Table 6.3: Ablation experiments for predicting Juul usage from tweets

| Model | F1 | 95% CI | Precision | 95% CI | Recall | 95% CI |
|---|---|---|---|---|---|---|
| **Wide&Deep** | **79.06%** | 78.95-79.17% | 80.61% | 80.52-80.69% | **77.58%** | 77.39-77.77% |
| - LSTM&Wide | 77.34% | 77.27-77.41% | 81.37% | 81.28-81.46% | 73.69% | 73.59-73.79% |
| - CNN&Wide | 58.45% | 58.24-58.67% | 63.01% | 62.87-63.16% | 54.54% | 54.18-54.89% |
| - LSTM | *77.85%* | 77.77-77.94% | **82.33%** | 82.20-82.45% | 73.84% | 73.73-73.96% |
| - CNN | 58.09% | 57.91-58.27% | 62.09% | 61.91-62.29% | 54.59% | 54.32-54.86% |
| - Wide | 77.48% | 77.33-77.63% | 81.59% | 81.45-81.72% | 73.78% | 73.52-74.04% |

## 6.3 Discussion

We choose a fine-tuned linear model with count vectorizer features as a baseline. The stacking model outperformed each traditional machine learning model. The TF-IDF feature vectorizer performed better than the count vectorizer, and the handcrafted features improve about 1% for count feature vectorizer and only about 0.2% for TF-IDF feature vectorizer. The possible reason is that the handcrafted features are count based, and hence are closer to what the count vectorizer is able to produce. Although the deep model and the stacking model with TF-IDF and handcrafted features have the same F1-score (italicized in Table 6.2), the deep model has higher precision than the stacking model. In this task, the precision has more top priority than the recall; therefore, the deep model outperforms the traditional machine learning approach. Overall, the best model is the Wide&Deep model with F1-score 79.06%, compared with the deep model, it has less than 1% decrease in precision, but almost 4% improvement in recall and the deep model alone has the best precision with 81.59%. In the ablation experiments, we found wide features are useful to CNN model, given deep model constructs features from scratch from training examples, and the wide features can provide prior knowledge of the domain. On the contrary, the LSTM with the wide features did not work as well; one possible reason is the wide feature contains a count of part-of-speech tags; this may bring a negative effect. The full model combining CNN, LSTM, and wide features improves recall by almost 4%. Our experiment on training Bi-LSTM on tweets also shows higher recall than precision. We think the CNN model leans more toward precision and LSTM model has better recall oriented setup.

In this section, we manually analyzed misclassified tweets. Since we did not impose a constraint on the length of the tweets, we compared the average length of the misclassified tweets (14.56 words) and the corresponding average for all test tweets (14.48 words), finding no significant difference. In the next step, we split the misclassified tweets into two groups of errors: type I (false positive) and type II (false negative). By manually examining the tweets of each group we had the following findings.

We found false positives in two different contexts. First, when there are first-person pronouns in the tweet and second when the tweet has quoted sentences representing potentially a conversation. Consider the following tweets:

- *Yeah, if the passenger in my car pulls out a juul i'm crashing idgaf*
- *hey everyone i stick my face in the mist from my humidifier and started coughing so idk how y'll are out here using a juul every 30 seconds*
- *This girl I work with she bought a Juul so she can attract friends when she goes places.*
- **jersey shore in 2018* "UBERS HERE" "Where the fks my Juul charger?!?" "Put that on the boardwalk snap story!" "Pla?*
- *"I think god wants me to juul because I have found two juuls on the ground" "god does not want that"*

The first three tweets contain first-person pronouns, but the keyword **juul** does not immediately follow them. A potential solution is to consider the distance between first-person pronouns and **juul**; it is possible to add the distance aspect as part of the wide model and/or create another LSTM model with the input of tweet content. The last two tweets in the list above have quotation marks to represent conversation. These two tweets contain first-person pronouns and also the distance between them and **juul** are closer than the previous three tweets. A solution here might involve regular expressions to find if the tweet is a conversation or not and add to the wide model.

Consider the following false negative tweets where we missed Juul user tweets:

- *I got a juul I can finally label myself as one of the cool kids*
- *buying a juul was the best decision i've made in the past 6 months*
- *If you are quitting juuling because of fake news I'll gladly take your pods*

To address these errors, we can use the attention mechanism to the deep model. The CNN model did not consider the word sequence information and cannot catch the relation between the word *juul* and pronoun *I*. The Bi-LSTM network consider the word sequence ;it can learn the difference between "I got a juul" and "juul I got". The word order can affect the prediction result. Therefore, using a Bi-LSTM network to replace CNN for the tweet text is an alternative way to consider.

### 6.4 Application of Juuler Model for Demographic Studies

In this section, we introduce a demographic study based on this juuler model. We used Twitter Streaming API to collect tweets that contain *juul, juuling, juuls* from 10/19/2018 to 6/8/2018 and got a total of 785,680 tweets. We followed our previous study (Han and Kavuluru, 2016; Kavuluru et al., 2019) to filter out the tweets which are created by an organization. By using Humanizr (McCorriston et al., 2015), we got 750,302 tweets that are classified as generated by an individual.

In this study, we want to estimate the percentage of Juul users in the teenager (age $<$ 18) and late adolescent (18 $<=$ age $<$ 21) groups. Since Twitter does not require users to disclose their age, we created an age model to estimate Twitter user's age. This age model was based on Zhang et al.'s (Zhang et al., 2016) effort. We trained two binary CNN models $C_1, C_2$, where $C_1$ is distinguishes between teens and adults, and $C_2$ classifies between late adolescents and adults. The model performance is shown in Table 6.4.

We combine the outputs from the age model and juuler model for further analysis. We found 127,601 Twitter users (out of 405,430 Twitter users who tweet about Juul)

Table 6.4: Age model performance

| Model | F1 | 95% CI | Precision | 95% CI | Recall | 95% CI |
|---|---|---|---|---|---|---|
| $C_1$ | 70.08% | 70.03-70.13% | 82.21% | 82.14-82.28% | 61.07% | 61.01-61.14% |
| $C_2$ | 74.55% | 74.49-74.61% | 86.17% | 86.11-86.23% | 65.69% | 65.61-65.78% |

are juulers. The teenager juuler count is 10,287, and the late adolescent juuler count is 23,891. To estimate the population distribution on Juul usage, we used Monte Carlo simulations (Mooney, 1997) to know the percentage of people using Juul.

Based on the U.S. Census 2016 (U.S. Census Bureau, 2017), we have age 18 and over constitute 76.90% of the U.S. population, and age 21 and over account for 72.70% of the population. From Pew Research (Pew Research Internet Project, 2013a) we know: 33% of teens use Twitter; 21% U.S. adults use Twitter; 36% online adults are between 18 and 29, and 31.5% of adults ages between 18 and 29 use Twitter. Therefore, we can calculate the online population based on this information, and the result is shown in Table 6.5. The baseline proportions of juuler and non-juuler distribution are 49.1% and 50.9%, respectively.

Table 6.5: Population distribution based on age

| | Teen (Age <18) | Late Adolescent (18–20) | Adult (21 and up) |
|---|---|---|---|
| US Population | 73,586,935 | 13,379,443 | 231,591,784 |
| Online Population | 24,283,689 | 4,816,599 | 46,627,358 |
| Percentage | 32.07% | 6.36% | 61.57% |

The distribution as per our age prediction model for Twitter users is 7.05%, 10.25%, and 82.7% for teenage, late adolescent, and adult populations, respectively. And Juuler distribution for Juul and non-Juul users is 31.47% and 68.53%, respectively, when we applied our Juul model on the full Twitter user dataset we collected. We simulate the combinations from age and Juul use groupings one million times and obtain the results in Table 6.6.

Table 6.6: Monto Carlo simulation results

| | Non-Juuler (Adult) | Non-Juuler (Teen/Adolescent) | Juuler(Adult) | Juuler (Teen/Adolescent) |
|---|---|---|---|---|
| Teen Baseline | 34.59% | 16.27% | 33.41% | 15.73% |
| Teen Predict | 63.68% | 4.86% | 29.25% | 2.21% |
| Adolescent Baseline | 31.33% | 19.63% | 30.17% | 18.87% |
| Adolescent Predict | 56.68% | 11.85% | 26.03% | 5.44% |

Comparing the simulation results, we can show that our conclusion is conservative. About 8,500 teenagers, and around 22,000 late adolescents in this dataset are Juul users. Since Juul's recommended minimum age is 21, our study shows at least 5% people didn't follow the recommendation.

## 6.5 Conclusion

In this chapter, we first estimated the proportion of the Juul user and non-user classes among Twitter users who posted messages about Juul. This was done by three annotators who assigned labels to a randomly selected set of 3000 tweets on Juul. Our analysis showed that class proportions are closed to balanced, which means almost 50% of people who were tweeting about Juul is a juuler. Based on the annotated dataset, we used supervised methods to create various classifiers with different features. Unlike prior efforts that employed a naive baseline without any fine-tuning, we created a strong baseline with fine-tuned hyperparameters shown in table 6.2. The primary measurement is F1-score with preference accorded for higher precision. The best model is our Wide&Deep ensemble with F1-score 79.06%, which shows about 2% improvement compared with the best traditional machine learning approach – stacking model. This study is classifying users whose tweets contain the term Juul. Therefore, the limitation of this study is that we cannot estimate the juuler proportion among all Twitter users. This model is useful as a fundamental step to conduct further demographic studies such as combining gender and age to analyze prevalence in different subpopulations. We can examine the attitudes of women and men towards Juul. We also performed a demographic study in Section 6.4 to show how to combine juuler model with other models for demographic attributes. These results may provide insights to different stakeholders in tobacco prevention including government agencies, non-profits, and policy makers in this controversial space of e-cigs.

In the future, we plan to improve our model by adding new handcrafted rules such as the distance between *personal pronouns* and the term *juul**, if the tweet is a conversation. In addition, we will also explore different deep neural network structures such as Bi-LSTM with an attention mechanisms. Using domain specific pre-trained embeddings and more recent language modeling based pre-trained networks could further help improve our models.

# Chapter 7 Conclusion and Future Work

Online health information and communication provides a new channel for patients, healthcare providers, the general public, and government agencies to glean information about diseases, provide online support and to assess consumer opinions. In the United States, more than 30% adults regularly use the internet to self-diagnose based on their symptoms. These e-patients usually start with online searches with Google to arrive at a self-diagnosis. However, there is no quality control in the online health information landscape. Sometimes the results can be confusing or even heavily misleading to the users. In the worst case, the patients may miss the critical period to obtain urgent care. Social media gives way for people to communicate on health topics. For example, an online mental support forum lets people communicate anonymously without disclosing personal identity. In addition, these support forums are usually moderated by trained professionals who can minimize the inaccurate information. Online health information is a broad space covering support forums, adverse drug reaction mentions, demographic studies on health-related issues, and policies. In this thesis, we applied text mining methods to answer various health-related questions. In the rest of this chapter, we talk about our contributions, limitations, and future works.

## 7.1 Summary of Contributions

In this dissertation, we used machine learning and deep learning approaches to solve three health-related questions. Our main contributions listed as follow:

**Incremental Under-sampling for Imbalanced Datasets.** In Chapter 4, we deal with a small and very imbalanced data set. The traditional approach for an imbalanced dataset includes over-sampling, under-sampling, and setting class weight. None of these approaches work well in this task. Therefore, instead of single time under-sampling, we proposed multiple under-sampling method with different class distributions. Based on the different distributions, we trained multiple models and averaged the prediction probability estimates to get the final prediction results. We show that incremental undersampling can better handle the imbalanced data, especially when the minority class is very small. Our approach tends to reduce the training sample noise arising from human labeling. Comparing with the traditional approach where different models are trained with the same dataset, our models are learned from a different subset in which the noisy data will have a smaller effect on the whole model.

**Attention based CNN for Adverse Drug Reaction Classification.** In Chapter 5, we presented a novel attention-based CNN for detecting ADR events. Instead of applied attention vector on the convolution layers as is typically done, we applied the attention vector on the max-pooling layers. Besides, we concatenated the BOW of full text to the max-pooling layer. The advantage of this architecture is that we teach models focusing on the most import n-grams, which are the terms related to the drugs and/or reaction events. Our result shows that this model improves by nearly 5% in F-score compared with the ensemble model between CNN only and the traditional machine learning approaches. The attention mechanism helps the model assess the relative discriminative power of different n-grams in the input.

**Wide and Deep model to Identify Juul Users.** In Chapter 6, we introduce our Wide and Deep model to identify Juul users among Twitter users who tweet about Juul. Our model uses the text inputs for CNN, part-of-the-speech tags as inputs for LSTM, and some handcraft features as wide features to build a state-of-the-art model. We found (from Chapter 6) the deep model has high precision, and the traditional machine learning approach is better from a recall perspective. With the Wide and Deep model, we took advantage of both deep and feature-engineered models to reach a new state-of-the-art result.

## 7.2 Limitations and Directions for Future Work

A general limitation of any social media based health research is the the relative differences in representation of different populations on online social networks and forums. The limitations arise from a distributional shift in general population and folks who post on social media. Social media based posts tend to come from more urban dwellers who are relatively younger. Even among youth, there are relative differences in which social network they use prominently. Literacy (education levels) and income groups also may affect what characterizes people who post on social streams. For our Juul study, we specifically focused on people who post about Juul while there could be many users who may not be posting messages on Juul. Then we used Monte Carlo simulation to estimate prevalence in the overall population. Thus, our result might under-estimate the real-world situation about underage Juul usage. Our Juul tweets analysis was based on 2018 data; with the most current news on adverse effects of e-cigarettes usage, the distribution of juuler and non-juuler on Twitter might change a lot. This change can affect our final estimation result.

In the future, we plan continue working in the following areas.

1. For triaging mental health messages, our results show co-training is helpful in the binary classification on green vs. non-green class. As we have a large number of unlabeled posts, we can continue to explore the model on how to use these large unlabeled messages to further improve scores.

2. For the adverse drug reaction detection task, more external domain knowledge is required. Therefore, using name entity recognition to capture the drug terms, UMLS as the source for synonyms, and MedDra for the adverse events can lead to future improvements compared with purely supervised methods used in our work thus far.

3. In this dissertation, we have focused on identifying Juul users, which can be treated as a fundamental step to do other demographic studies. For instance, we can look different segments of population by using models that predict gender, income range, education level, and race besides the age model we used in our efforts. We can further extend this by looking at side affects experienced by Juul users as disclosed on social networks.

## Abbreviations

**ADR** Adverse Drug Reaction. 47, 48, 59

**AMIA** American Medical Informatics Association. 47

**Bi-LSTM** Bidirectional-Long Short Term Memory. 66–68, 70

**CF** Convolutional Filter. 39, 50

**CLPsych** Computational Linguistics and Clinical Psychology Workshop. 34

**CNN** Convolutional Neural Network. 14, 40, 47, 50, 63, 66, 72

**CNN-Att** Convolutional Neural Network with Attention. 52, 54

**CSN** Cancer Survivors Network. 8

**DNN** Deep Neural Network. 14, 39

**E-cigs** Electronic cigarettes. 58

**FDA** Food and Drug Administration. 47

**GBRT** Gradient Boosting Regression Trees. 58

**GPU** Graphics Processing Unit. 15

**IAA** Inter-Annotator Agreement. 47, 55

**k-NN** k-Nearest Neighbors. 38, 43

**LIWC** Linguistic Inquiry and Word Count. 35, 36

**LR** Logistic Regression. 38, 43

**LSTM** Long Short Term Memory. 39, 40, 60, 66, 67, 72

**MLP** Multilayer Perceptron. 13

**NER** Named Entity Recognition. 10

**NLP** Nature Language Processing. 2

**NRC** National Research Council. 35

**OHCs** Online Health Communities. 3

**PMI** Pointwise Mutual Information. 48, 49

**RBF** Radial Basis Function. 34, 38

**ReHAN** Recurrent neural networks with Hierarchical Attention. 59

**RF** Random Forests. 64

**SGD** Stochastic Gradient Descent. 34

**SIDS** Sudden Infant Death Syndrome. 7

**SMM4h** Social Media Mining for Health. 5

**SVM** Support Vector Machine. 12, 34, 38, 43, 47

**TF-IDF** Term Frequency-Inverse Document Frequency. 35, 66

**WHO** World Health Organization. 33

**XML** eXtensible Markup Language. 34

# Bibliography

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. "Sentiment analysis of twitter data". In: *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics. 2011, pp. 30–38.

[2] I. Alexa. *Alexa top 500 global sites*. `http://www.alexa.com/topsites`. 2016.

[3] E. Aramaki, S. Maskawa, and M. Morita. "Twitter catches the flu: detecting influenza epidemics using Twitter". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, pp. 1568–1576.

[4] D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[5] J. L. Barrington-Trimis, R. Urman, K. Berhane, J. B. Unger, T. B. Cruz, M. A. Pentz, J. M. Samet, A. M. Leventhal, and R. McConnell. "E-Cigarettes and Future Cigarette Use". In: *Pediatrics* (2016), e20160379.

[6] D. M. Blei and J. D. Lafferty. "Topic Models". In: *Text Mining: Classification, Clustering, and Applications*. Ed. by A. Srivastava and M. Sahami. Chapman and Hall, CRC Press, 2009. Chap. 4, pp. 71–93.

[7] A. Blum and T. Mitchell. "Combining labeled and unlabeled data with co-training". In: *Proceedings of the eleventh annual conference on Computational learning theory*. ACM. 1998, pp. 92–100.

[8] G. Bouma. "Normalized (pointwise) mutual information in collocation extraction". In: *Proceedings of GSCL* (2009), pp. 31–40.

[9] C. Brew. "Classifying ReachOut posts with a radial basis function SVM". In: *Proceedings of the third workshop on computational lingusitics and clinical psychology*. 2016, pp. 138–142.

[10] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. "Digital disease detection—harnessing the Web for public health surveillance". In: *New England Journal of Medicine* 360.21 (2009), pp. 2153–2157.

[11] Centers for Disease Control. *E-cigarette use triples among middle and high school students in just one year*. 2015.

[12] B. P. Chamberlain, C. Humby, and M. P. Deisenroth. “Probabilistic inference of twitter users’ age based on what they follow”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2017, pp. 191–203.

[13] A. J.-B. Chaney and D. M. Blei. “Visualizing Topic Models”. In: *International Conference of Weblogs and Social Media*. ICWSM ’12. 2012.

[14] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen. “Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 3036–3044.

[15] I.-L. Chen et al. “FDA summary of adverse events on electronic cigarettes”. In: *Nicotine & Tobacco Research* 15.2 (2013), pp. 615–616.

[16] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. “Wide & deep learning for recommender systems”. In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. ACM. 2016, pp. 7–10.

[17] X. Cheng, X. Yan, Y. Lan, and J. Guo. “BTM: Topic modeling over short texts”. In: *Knowledge and Data Engineering, IEEE Transactions on* 26.12 (2014), pp. 2928–2941.

[18] J. Choi, Y. Cho, E. Shim, and H. Woo. “Web-based infectious disease surveillance systems and public health perspectives: a systematic review”. In: *BMC public health* 16.1 (2016), p. 1238.

[19] W.-Y. S. Chou, Y. M. Hunt, E. B. Beckjord, R. P. Moser, and B. W. Hesse. “Social media use in the United States: implications for health communication”. In: *Journal of medical Internet research* 11.4 (2009).

[20] K.-H. Chu, J. B. Unger, J.-P. Allem, M. Pattarroyo, D. Soto, T. B. Cruz, H. Yang, L. Jiang, and C. C. Yang. “Diffusion of Messages from an Electronic Cigarette Brand to Potential Users through Twitter”. In: *PloS one* 10.12 (2015), e0145387.

[21] M. Chung, R. P. Oden, B. L. Joyner, A. Sims, and R. Y. Moon. “Safe infant sleep recommendations on the Internet: let’s Google it”. In: *The Journal of pediatrics* 161.6 (2012), pp. 1080–1084.

[22] T. Cohen, D. Widdows, L. De Vine, R. Schvaneveldt, and T. C. Rindflesch. "Many paths lead to discovery: analogical retrieval of cancer therapies". In: *Quantum Interaction*. Springer, 2012, pp. 90–101.

[23] H. Cole-Lewis, J. Pugatch, A. Sanders, A. Varghese, S. Posada, C. Yun, M. Schwarz, and E. Augustson. "Social Listening: A Content Analysis of E-Cigarette Discussions on Twitter". In: *Journal of medical Internet research* 17.10 (2015).

[24] H. Cole-Lewis, A. Varghese, A. Sanders, M. Schwarz, J. Pugatch, and E. Augustson. "Assessing electronic cigarette-related Tweets for sentiment and content using supervised machine learning". In: *J. of medical Internet research* 17.8 (2015), e208.

[25] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. "Natural language processing (almost) from scratch". In: *Journal of machine learning research* 12.Aug (2011), pp. 2493–2537.

[26] A. Culotta, N. R. Kumar, and J. Cutler. "Predicting the Demographics of Twitter Users from Website Traffic Data". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015, pp. 72–78.

[27] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, et al. "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality". In: *Psychological science* 26.2 (2015), pp. 159–169.

[28] J.-F. Etter, C. Bullen, A. D. Flouris, M. Laugesen, and T. Eissenberg. "Electronic nicotine delivery systems: a research agenda". In: *Tobacco Control* 20.3 (2011), pp. 243–248.

[29] Food and Drug Administration, HHS et al. "Deeming Tobacco Products To Be Subject to the Federal Food, Drug, and Cosmetic Act, as Amended by the Family Smoking Prevention and Tobacco Control Act; Restrictions on the Sale and Distribution of Tobacco Products and Required Warning Statements for Tobacco Products. Final rule." In: *Federal register* 81.90 (2016), p. 28973.

[30] A. K. Godea, C. Caragea, F. A. Bulgarov, and S. Ramisetty-Mikler. "An Analysis of Twitter Data on E-cigarette Sentiments and Promotion". In: *Conference on Artificial Intelligence in Medicine in Europe*. Springer. 2015, pp. 205–215.

[31] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle. "Multimedia lab@ acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations". In: *ACL-IJCNLP* 2015 (2015), pp. 146–153.

[32] N. A. Goriounova and H. D. Mansvelder. "Short-and long-term consequences of nicotine exposure during adolescence for prefrontal cortex neuronal network function". In: *Cold Spring Harbor perspectives in medicine* (2012), a012120.

[33] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[34] S. Han and R. Kavuluru. "Exploratory analysis of marketing and non-marketing e-cigarette themes on Twitter". In: *International Conference on Social Informatics*. Springer. 2016, pp. 307–322.

[35] S. Han and R. Kavuluru. "On Assessing the Sentiment of General Tweets". In: *Canadian Conference on Artificial Intelligence*. Springer. 2015, pp. 181–195.

[36] S. Han, T. Tran, A. Rios, and R. Kavuluru. "Team UKNLP: Detecting ADRs, Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter." In: *SMM4H@ AMIA*. 2017, pp. 49–53.

[37] H. M. Harb and A. S. Desuky. "Feature selection on classification of medical datasets based on particle swarm optimization". In: *International Journal of Computer Applications* 104.5 (2014).

[38] B. Herzog and P. Kanada. *Nielsen: Tobacco "all channel" data 1/27*. Feb. 2018. URL: https://1lbxcx1bcuig1rfxaq3rd6w9-wpengine.netdna-ssl.com/wp-content/uploads/2018/02/Nielsen-Tobacco-All-Channel-Report-Period-Ending-1.27.18.pdf.

[39] M. Hoffman, F. R. Bach, and D. M. Blei. "Online learning for latent Dirichlet allocation". In: *Advances in neural information proc. systems*. 2010, pp. 856–864.

[40] L. Hong and B. D. Davison. "Empirical study of topic modeling in twitter". In: *Proc. of the 1st workshop on social media analytics*. ACM. 2010, pp. 80–88.

[41] M. Honnibal and M. Johnson. "An Improved Non-monotonic Transition System for Dependency Parsing". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1373–1378. URL: https://aclweb.org/anthology/D/D15/D15-1162.

[42] J. Huang, R. Kornfield, G. Szczypka, and S. L. Emery. “A cross-sectional examination of marketing of electronic cigarettes on Twitter”. In: *Tobacco control* 23.suppl 3 (2014), pp. iii26–iii30.

[43] D. M. E.-D. M. Hussein. “A survey on sentiment analysis challenges”. In: *Journal of King Saud University-Engineering Sciences* (2016).

[44] T. Initiative. *6 important facts about JUUL*. Aug. 2018. URL: `https://truthinitiative.org/news/6-important-facts-about-juul`.

[45] T. Initiative. *New study: Teens 16x more likely to use JUUL than older age groups*. Nov. 2018. URL: `https://truthinitiative.org/news/new-study-reveals-teens-16-times-more-likely-use-juul-older-age-groups`.

[46] S. Ioffe and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015).

[47] D. Jain and V. Singh. “Feature selection and classification systems for chronic disease prediction: A review”. In: *Egyptian Informatics Journal* 19.3 (2018), pp. 179–189.

[48] R. Kavuluru, S. Han, and E. J. Hahn. “On the popularity of the USB flash drive-shaped electronic cigarette Juul”. In: *Tobacco control* 28.1 (2019), pp. 110–112.

[49] R. Kavuluru, A. Rios, and T. Tran. “Extracting Drug-Drug Interactions with Word and Character-Level Recurrent Neural Networks”. In: *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*. IEEE. 2017, pp. 5–12.

[50] R. Kavuluru and A. Sabbir. “Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter”. In: *J. of biomedical informatics* 61 (2016), pp. 19–26.

[51] R. Kavuluru, A. G. Williams, M. Ramos-Morales, L. Haye, T. Holaday, and J. Cerel. “Classification of helpful comments on online suicide watch forums”. In: *ACM-BCB......: the... ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. Vol. 2016. NIH Public Access. 2016, p. 32.

[52] A. E. Kim, T. Hopper, S. Simpson, J. Nonnemaker, A. J. Lieberman, H. Hansen, J. Guillory, and L. Porter. "Using Twitter Data to Gain Insights into E-cigarette Marketing and Locations of Use: An Infoveillance Study". In: *Journal of Medical Internet Research* 17.11 (2015), e251.

[53] A. Kim, T. Miano, R. Chew, M. Eggers, and J. Nonnemaker. "Classification of Twitter users who tweet about e-cigarettes". In: *JMIR public health and surveillance* 3.3 (2017).

[54] Y. Kim. "Convolutional Neural Networks for Sentence Classification". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1746–1751.

[55] Y. Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

[56] E. G. Klein, M. Berman, N. Hemmerich, C. Carlson, S. Htut, and M. Slater. "Online E-cigarette Marketing Claims: A Systematic Content and Legal Analysis". In: *Tobacco Regulatory Science* 2.3 (2016), pp. 252–262.

[57] J. Landis and G. Koch. "The measurement of observer agreement for categorical data." In: *Biometrics* 33.1 (1977), pp. 159–174.

[58] D. T. Levy, K. M. Cummings, A. C. Villanti, R. Niaura, D. B. Abrams, G. T. Fong, and R. Borland. "A framework for evaluating the public health impact of e-cigarettes and other vaporized nicotine products". In: *Addiction* (2016).

[59] L. Li, L. Jin, Z. Jiang, D. Song, and D. Huang. "Biomedical named entity recognition based on extended recurrent neural networks". In: *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE. 2015, pp. 649–652.

[60] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. "Automatic discovery and optimization of parts for image classification". In: *International Conference on Learning Representations (ICLR)*. 2017.

[61] W. Liu and D. Ruths. "What's in a Name? Using First Names as Features for Gender Inference in Twitter." In: *Proceedings of the AAAI Spring Symposium: Analyzing Microtext*. 2013, pp. 10–16.

[62] D. D. Luxton, J. D. June, and J. M. Fairall. "Social media and suicide: a public health perspective". In: *American journal of public health* 102.S2 (2012), S195–S200.

[63] S. Mac Kim, Y. Wang, S. Wan, and C. Paris. “Data61-csiro systems at the clpsych 2016 shared task”. In: *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. 2016, pp. 128–132.

[64] S. Malik, A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, and B. Shneiderman. “Topicflow: visualizing topic alignment of twitter data over time”. In: *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*. ACM. 2013, pp. 720–726.

[65] S. Malmasi, M. Zampieri, and M. Dras. “Predicting post severity in mental health forums”. In: *Proceedings of the third workshop on computational lingusitics and clinical psychology*. 2016, pp. 133–137.

[66] D. Mangan. *E-cigarette sales are smoking hot at $1.7 billion*. Dec. 2015. URL: `https://www.cnbc.com/id/100991511`.

[67] E. Martin, P. W. Clapp, M. E. Rebuli, E. A. Pawlak, E. E. Glista-Baker, N. L. Benowitz, R. C. Fry, and I. Jaspers. “E-cigarette use results in suppression of immune and inflammatory-response genes in nasal epithelial cells similar to cigarette smoke”. In: *American Journal of Physiology-Lung Cellular and Molecular Physiology* (2016), ajplung–00170.

[68] J. McCorriston, D. Jurgens, and D. Ruths. “Organizations are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter”. In: *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2015.

[69] A. McNeill, L. Brose, R. Calder, S. Hitchman, P. Hajek, and H. McRobbie. “E-cigarettes: an evidence update”. In: *Report from Public Health England* (2015).

[70] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. “Distributed representations of words and phrases and their compositionality”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2013, pp. 3111–3119.

[71] S. M. Mohammad, S. Kiritchenko, and X. Zhu. “NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets”. In: *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*. Atlanta, Georgia, USA, June 2013.

[72] C. Z. Mooney. *Monte carlo simulation*. Vol. 116. Sage Publications, 1997.

[73] M. Myslin, S.-H. Zhu, W. Chapman, and M. Conway. "Using twitter to examine smoking behavior and perceptions of emerging tobacco products". In: *Journal of medical Internet research* 15.8 (2013).

[74] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. "Semeval-2013 task 2: Sentiment analysis in twitter". In: *Proc. SemEval* (2013).

[75] NAMI. *Mental Health By the Numbers.* 2019. URL: `https://www.nami.org/learn-more/mental-health-by-the-numbers`.

[76] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. ""How Old Do You Think I Am?" A Study of Language and Age in Twitter." In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM).* 2013, pp. 439–448.

[77] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham. "An analysis of the coherence of descriptors in topic modeling". In: *Expert Systems with Applications* 42.13 (2015), pp. 5645–5657.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[81] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew. "Sentiment analysis of suicide notes: A shared task". In: *Biomedical informatics insights* 5.Suppl 1 (2012), p. 3.

[82] Pew Research Internet Project. *Part 1: Teens and Social Media Use.* 2013.

[83] Pew Research Internet Project. *Part 1: Teens and social media use.* http://www.pewinternet.org/2013/05/21/part-1-teens-and-social-media-use/. 2013.

[84] Pew Research Internet Project. *Social media update.* http://www.pewinternet.org/2013/12/30/social-media-update-2013/. 2013.

[85] Pew Research Internet Project. *Social media usage: 2005-2015.* 2015.

[86] Piper Jaffray Market Research Project. *Taking stock with teens.* http://www.piperjaffray.com/3col.aspx?id=3045. 2014.

[87] C. Quan, L. Hua, X. Sun, and W. Bai. "Multichannel convolutional neural network for biological relation extraction". In: *BioMed research international* 2016 (2016).

[88] S. Raschka. "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack". In: *The Journal of Open Source Software* 3.24 (Apr. 2018). DOI: 10.21105/joss.00638. URL: http://joss.theoj.org/papers/10.21105/joss.00638.

[89] R. Řehůřek and P. Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[90] A. Rios and R. Kavuluru. "Convolutional neural networks for biomedical text classification: application in indexing biomedical articles". In: *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics.* ACM. 2015, pp. 258–267.

[91] S. Rudy and E. Durmowicz. "Electronic nicotine delivery systems: overheating, fires and explosions." In: *Tobacco control* (2016).

[92] A. Sarker and G. Gonzalez. "Portable automatic text classification for adverse drug reaction detection via multi-corpus training". In: *Journal of biomedical informatics* 53 (2015), pp. 196–207.

[93] S. Sasikala, S. A. alias Balamurugan, and S. Geetha. "Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set". In: *Applied Computing and Informatics* 12.2 (2016), pp. 117–127.

[94] T. Singh, R. Arrazola, C. Corey, C. Husten, L. Neff, D. Homa, and B. King. "Tobacco Use Among Middle and High School Students – United States, 2011 – 2015." In: *MMWR Morbidity and mortality weekly report* 65.14 (2016), pp. 361–367.

[95] J. Sultana, P. Cutroneo, and G. Trifirò. "Clinical and economic burden of adverse drug reactions". In: *Journal of pharmacology & pharmacotherapeutics* 4.Suppl1 (2013), S73.

[96] M. Sundermeyer, R. Schlüter, and H. Ney. "LSTM neural networks for language modeling". In: *Thirteenth annual conference of the international speech communication association*. 2012.

[97] Y. R. Tausczik and J. W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54.

[98] T. Tran and R. Kavuluru. "Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks". In: *Journal of biomedical informatics* 75 (2017), S138–S148.

[99] Twitter, Inc. *Registration with United States Securities and Exchanges Commission*. 2013.

[100] P. D. U.S. Census Bureau. *Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States, States, Counties and Puerto Rico Commonwealth and Municipios: April 1, 2010 to July 1, 2016*. `https://factfinder.census.gov`. June 2017.

[101] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. "Class imbalance, redux". In: *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE. 2011, pp. 754–763.

[102] *What is Health Communications?* `https://www.cdc.gov/healthcommunication/healthbasics/whatishc.html`. 2004.

[103] P. Wicks, M. Massagli, J. Frost, C. Brownstein, S. Okun, T. Vaughan, R. Bradley, and J. Heywood. "Sharing health data for better outcomes on PatientsLikeMe". In: *Journal of medical Internet research* 12.2 (2010).

[104] E. B. Wilson. "Probable inference, the law of succession, and statistical inference". In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212.

[105] J. Zhang, X. Hu, Y. Zhang, and H. Liu. “Your age is no secret: Inferring microbloggers’ ages via content and interaction analysis”. In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[106] M. Zhu, C. Xu, and Y.-F. B. Wu. “IFME: information filtering by multiple examples with under-sampling in a digital library environment”. In: *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM. 2013, pp. 107–110.

## Vita

### Name

Sifei Han

### Education

- 2008–2012 B.S. in Computer Science University of Kentucky Lexington, Kentucky
- 2008–2012 B.S. in Mathematical Economics University of Kentucky Lexington, Kentucky

### Experience

- 2012–present, Graduate Research Assistant, University of Kentucky, Lexington, Kentucky
- 2013–2017, Graduate Teaching Assistant, University of Kentucky, Lexington, Kentucky
- 2016, Software Engineer, Office of Sponsored Projects Administration at University of Kentucky, Lexington, Kentucky
- 2012, SharePoint Designer Intern, AllTech Inc., Nicholasville, Kentucky
- 2011, Undergraduate Teaching Assistant, University of Kentucky, Lexington, Kentucky

### Awards

- 2017 – Ranked 2nd (among 11 teams) in the shared task on classification of medication intake messages on Twitter for online pharmacovigilance (at Social media mining for health workshop at AMIA)
- 2016 – Omicron Delta Kappa (ODK) national leadership honor society
- 2015 – University of Kentucky Graduate Student Travel Grant.
- 2013 – University of Kentucky Graduate Student Travel Grant.

- 2011 – Dean’s List, University of Kentucky
- 2010 – Strunk Thurston H. Scholarship, University of Kentucky
- 2008 – Flagship Scholarship, University of Kentucky.

**Publications**

1. A. Sarker, M. Belousov, J. Friedrichs, K. Hakala, S. Kiritchenko, F Mehryary, **S Han**, T. Tran, A. Rios, R. Kavuluru, B. de Bruijn, F. Ginter4, D. Mahata, S. M. Mohammad, G. Nenadic, G. Gonzalez-Hernandez. Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H) 2017 shared task. JAMIA. 2018.
2. R. Kavuluru and **S. Han** and E. Hahn. On the popularity of the USB flash drive-shaped electronic cigarette Juul. Tobacco Control. 2018
3. **S.Han**, T. Tran, A.Rios,R.Kavuluru. Team UKNLP:Detecting ADRs,Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter. In Proceedings of the 2nd Social Media Mining for Health Applications Workshop and Shared Task at AMIA, 2017.
4. **S.Han**, and R.Kavuluru. "Exploratory analysis of marketing and non-marketing e-cigarette themes on Twitter." International Conference on Social Informatics. Springer International Publishing, 2016.
5. **S.Han**, and R.Kavuluru. "On assessing the sentiment of general tweets." Canadian Conference on Artificial Intelligence. Springer, Cham, 2015.
6. R.Kavuluru, **S.Han**, and D.Harris. "Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques." Canadian Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2013.