Journal of the American Medical Informatics Association, 25(1), 2018, 72–80
doi: 10.1093/jamia/ocx045
Advance Access Publication Date: 13 May 2017
Research and Applications

Research and Applications

Mining e-cigarette adverse events in social media using Bi-LSTM recurrent neural network with word embedding representation

Jiaheng Xie,[1] Xiao Liu,[2] and Daniel Dajun Zeng[1,3]

[1]Department of Management Information Systems, University of Arizona, Tucson, AZ, USA, [2]Department of Operation and Information Systems, University of Utah, Salt Lake City, UT, USA and [3]State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Corresponding Author: Jiaheng Xie, Department of Management Information Systems, 1130 E Helen St, McClelland Hall 430, Tucson, AZ 85721-0108, USA. E-mail: xiej@email.arizona.edu. Phone: +1 521-621-2748

Received 17 September 2016; Revised 3 April 2017; Accepted 11 April 2017

ABSTRACT

Objective: Recent years have seen increased worldwide popularity of e-cigarette use. However, the risks of e-cigarettes are underexamined. Most e-cigarette adverse event studies have achieved low detection rates due to limited subject sample sizes in the experiments and surveys. Social media provides a large data repository of consumers' e-cigarette feedback and experiences, which are useful for e-cigarette safety surveillance. However, it is difficult to automatically interpret the informal and nontechnical consumer vocabulary about e-cigarettes in social media. This issue hinders the use of social media content for e-cigarette safety surveillance. Recent developments in deep neural network methods have shown promise for named entity extraction from noisy text. Motivated by these observations, we aimed to design a deep neural network approach to extract e-cigarette safety information in social media.

Methods: Our deep neural language model utilizes word embedding as the representation of text input and recognizes named entity types with the state-of-the-art Bidirectional Long Short-Term Memory (Bi-LSTM) Recurrent Neural Network.

Results: Our Bi-LSTM model achieved the best performance compared to 3 baseline models, with a precision of 94.10%, a recall of 91.80%, and an F-measure of 92.94%. We identified 1591 unique adverse events and 9930 unique e-cigarette components (ie, chemicals, flavors, and devices) from our research testbed.

Conclusion: Although the conditional random field baseline model had slightly better precision than our approach, our Bi-LSTM model achieved much higher recall, resulting in the best F-measure. Our method can be generalized to extract medical concepts from social media for other medical applications.

Key words: E-cigarette adverse event, Bi-LSTM, recurrent neural network, word embedding, deep neural network

INTRODUCTION

An electronic cigarette, or e-cigarette, is an electronic nicotine delivery system (ENDS) that delivers a heated aerosol of nicotine in a fashion that mimics conventional cigarettes.[1] E-cigarettes have grown in popularity among all age groups. Research shows that 12.6% of adults in the United States have tried e-cigarettes in their lifetime.[2] E-cigarettes have a large market among youth as well.[3] According to the US Food and Drug Administration (FDA), e-cigarette use among high school students surged from 1.5% in 2011 to 16% in 2015.[3] The global e-cigarette industry is expected to grow over 22.36% from 2015 to 2025, reaching a total market value of $50 billion.[4]

Medical studies have discovered many e-cigarette adverse events that affect large groups of users.[5,6] Recognizing the potential risks

© The Author 2017. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

1) I$_{People}$ too only vape PG$_{Chemical}$ free liquid and do find if I$_{People}$ do not keep my$_{People}$ hydration$_{Event}$ levels up, I$_{People}$ tend to notice certain things like headaches$_{Event}$ the net morning$_{Time}$.
2) On the times I$_{People}$ tried to vape around 10$_{Number}$ Watts$_{Unit}$ with 16-18mg$_{Number}$, it was so hard I$_{People}$ need to stop, the nicotine$_{Chemical}$ rush was so big I$_{People}$ was nauseous$_{Event}$ and shaking.

Figure 1. Examples of e-cigarette discussions in social media

and the growing popularity of e-cigarette use, the FDA issued new regulations on ENDS in May 2016. The new rules mandate that e-cigarette manufacturers evaluate the ingredients of tobacco products and communicate their potential risks.[7] However, most e-cigarette manufacturers have failed to do so.[8] To detect hazards related to e-cigarettes and alert the public, regulatory agencies and public health researchers need to take a more proactive role.

Most existing studies on e-cigarettes developed clinical trials to evaluate the safety issues, with limited results due to their small sample size and short duration. Thus, the safety profile of e-cigarettes based on clinical trials is incomplete. It is essential to conduct postmarket surveillance and monitor the hazards related to e-cigarettes. Although postmarket monitoring systems, such as the FDA's MedWatch, have been established to collect reports of adverse events, most users are not aware of such systems. Consequently, a significant number of e-cigarette adverse events have never been reported. Also, current methods to detect e-cigarette adverse events, such as clinical experiments and the MedWatch system, require huge human capital and research budgets. There is a great need to explore cost-efficient reporting channels for e-cigarette adverse events.

As social media has grown in popularity, many e-cigarette discussion forums have emerged and received attention from vendors, consumers, health care professionals, and other stakeholders. Their social media engagements have become a valuable source for understanding e-cigarette user behavior, health effects, and marketing practices. Figure 1 shows 2 examples of e-cigarette discussions in social media.

Recognizing the value of consumer-generated content about e-cigarettes in social media, we aimed to develop a natural language processing approach to understanding e-cigarette safety issues. Our research testbed is composed of posts from the world's largest e-cigarette discussion forum. Leveraging recent developments in deep neural network methods, we utilized word embedding to represent semantic meaning in consumer vocabulary. We developed a novel Bidirectional Long Short-Term Memory (Bi-LSTM) Recurrent Neural Network model to identify e-cigarette safety issues from unstructured social media text. To the best of our knowledge, this study is among the first to develop a deep neural network approach for medical entity recognition in social media. Our study contributes to both health information extraction methodology and regulatory practice for ENDS product safety surveillance. By incorporating the LSTM unit and the bidirectional architecture, our Bi-LSTM model can be generalized to various entity recognition problems in the health informatics domain. The adverse events identified in this study can be referenced by regulatory agencies for decision support.

LITERATURE REVIEW

E-cigarette adverse events

E-cigarette safety monitoring usually has 2 phases: premarketing review and postmarketing surveillance. In the premarketing phase, e-cigarette vendors,[9] the FDA,[10] and researchers[11–13] rely on experiments or surveys to examine e-cigarette safety with a small sample of users (10–30 subjects).[14–17] These surveys and experiments have identified adverse events such as increased heart rate,[16] decreased fractional exhaled nitric oxide,[17] increased white blood cells,[14] and increased interleukins and epidermal growth factor.[15] However, due to limited sample size and short experiment duration, many review studies are incapable of detecting rare adverse events and long-term effects of e-cigarettes.[18,19] Moreover, most premarketing e-cigarette review studies are conducted in controlled settings with constrained use cases. Risks associated with extreme cases such as high temperature are rarely examined. As a result, many adverse events that could affect a massive consumer base cannot be detected in premarketing reviews. To this end, large-scale postmarketing surveillance is essential for building safety profiles of e-cigarette products.

The responsibility for current postmarketing surveillance of e-cigarettes mainly lies in the FDA's MedWatch system. The MedWatch system allows users to report adverse events for medical products.[20] However, this voluntary reporting system has not shown success in e-cigarette safety monitoring. Since MedWatch started to collect reports about e-cigarette safety issues in 2008, the annual report numbers have not exceeded 30.[6]

While regulatory agencies struggle to obtain information about e-cigarettes, consumers have shared a significant amount of their experience on social media, making it a promising source for collecting reports of e-cigarette adverse events. For instance, E-Cigarette Forum, the largest social media platform for e-cigarette consumers in the world, contains over 17 million posts from about 250 000 registered members. To utilize social media data for e-cigarette adverse event detection, we need to identify entity names such as chemical compounds and medical events.[21] However, very few studies on e-cigarette safety monitoring have utilized social media data. To understand the state-of-the-art techniques of medical entity recognition in social media, we reviewed social media surveillance studies for medical products.

Adverse event extraction in social media

Social media has been adopted for postmarket surveillance of many medical products.[22–26] These studies demonstrate the value of social media in complementing the traditional medical product safety monitoring systems. Two common approaches have been developed for medical entity extraction: lexicon-based and statistical learning methods. Lexicon-based approaches leverage medical term dictionaries, such as the Unified Medical Language System (UMLS),[27] GATE,[28] and MedLEE,[29] to map user-generated words to standard medical concepts and their entity types. They usually achieve low accuracy (around 20–60%),[22,30,31] because lexicon-based methods cannot detect variations of medical terms used in social media, such as consumer vocabularies, typos, and abbreviations. The state-of-the-art statistical learning method, conditional random fields (CRFs), usually achieve a higher precision (60–90%) than lexicon-based methods.[32–34] Nevertheless, their recalls are still low (only 40–70%),[35,36] because CRFs treat words as discrete atomic symbols and require accurate input for training and prediction. Unfortunately, in social media, there are many word variations, and hence

Table 1. Entity types

Entity Type	Explanation	Examples
Body part	Part of the human body	Chest, head, throat
Chem	Chemical compounds and components of e-liquid (chemical liquid used in e-cigarettes)	VG, PG, nicotine
Device	E-cigarette device	Madvapes, anjelvape77, atomizer
Event	User-reported adverse events after using e-cigarette	Headache, vomit, dizzy
Flavor	E-liquid flavors	Banana, cherry, apple
Num	Numbers	10, 20 mg, 5.23
People	People	I, son, Peter
Time	Time	Today, 2015, 3:50
Unit	Unit of weight, distance, etc.	Watts, mg, mile
Na	Any other mentioned entities	Have, only, on, around

CRFs are not as successful as they are in extracting entities from other text genres. More advanced models that can address the word variations and data sparsity issues are needed.

Recent developments in deep neural networks address the word variation issue by capturing the semantic meanings of words and achieve higher performance in entity extraction.[37] According to the distributional hypothesis, words with similar meanings occur with similar neighbors.[38] Word embedding represents each word in a vector of its surrounding words. Such a method enables us to represent a medical term with its semantic context instead of the symbolic term itself. We can effectively represent sparse entities, entities with typos, and entities with variations in social media data using word embedding. A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a loop. This loop creates an internal state in the network and enables information to persist during the learning process.[39] RNN models with word embedding input achieve good performance in many sequence learning tasks, such as part-of-speech tagging,[39] named entity recognition (NER),[40] and machine translation.[41] However, a standard RNN is not capable of learning long-term dependency (long input sequence),[42] as is often seen in social media text. Long Short-Term Memory (LSTM) RNNs, an improvement over standard RNNs, address the long-term dependency issue. An LSTM RNN can add or remove information in each internal state through the internal gates in the LSTM units. Both standard RNNs and LSTM RNNs have restrictions, as future input information cannot be reached at the current state. A Bi-LSTM RNN connects 2 hidden layers from opposite directions to the same output, thus making future input data available for the current state. Bi-LSTM RNNs have achieved leading performance for NER on noisy user-generated text[37,40] because of their ability to consider interconnected information in a sentence.

Research gaps

Although social media provides information related to e-cigarette safety on a large scale, there is very limited research in this area. The primary challenge is to interpret consumer health vocabulary (CHV) and extract useful information for e-cigarette safety monitoring from user-generated textual data. To address this issue, we propose to develop a Bi-LSTM model for e-cigarette adverse event detection. Our approach incorporates state-of-the-art word embedding and Bi-LSTM to identify e-cigarette components and medical events from social media. Word embedding can address the sparse entity and word variation issues in social media. The Bi-LSTM model can map an input embedding sequence to the predefined entity types with high performance. Our proposed method can also be utilized to solve various social media mining problems, such as adverse drug event detection, drug-drug interactions, and more.

METHODS

Data collection and annotation

Our research data was collected from E-Cigarette Forum (https://www.e-cigarette-forum.com/forum/), the largest e-cigarette online forum in the world. To ensure that the discussions were relevant to e-cigarette safety issues, we filtered the posts by focusing our analysis on 3 subforums: Health, Safety, and Vaping (vaping is the behavior of inhaling and exhaling vapor generated by e-cigarette devices); Tobacco Harm Reduction; and Nicotine. These 3 subforums were selected because their discussions mostly concentrate on e-cigarette safety issues such as adverse events. Our testbed encompasses 197 106 users, 6 054 832 posts, 155 296 threads, and 64 e-cigarette brands. We collected posts from April 1, 2008, to September 9, 2015.

Since 1 post can contain multiple sentences, we segmented the posts into sentences with the sentence boundary detection package from NLTK (http://www.nltk.org). Five thousand sentences were randomly selected from the testbed, 4000 in the training set and 1000 in the test set. Two expert annotators independently labeled the sentences for entity types. We list 10 entity types included in our annotation in Table 1.

Each word was labeled with an entity type. Figure 2 shows an example of sentence annotation; "eos" stands for "end of the sentence."

To measure interannotator reliability, we used Cohen's kappa.[43] The kappa value is 0.96 for the e-cigarette forum data annotation. A third annotator reviewed the disagreements and made the final judgment. Finally, the ground truth was generated, containing 4000 training sentences and 1000 test sentences. The statistics of the training and test sets are shown in Table 2.

Word embedding representation

We first trained an embedding model using the Skip-gram method in Word2vec[44] on the entire forum corpus. Each unique word in the corpus was assigned a number. The Skip-gram method predicts what words are likely to co-occur with the word of interest. The formula and description below detail the model. T is the number of unique words in the training corpus; c is the window size of the surrounding words. Words that appear within the distant of c are considered as the surrounding words. w_{t+j} are the words surrounding w_t. Given a word w_t, the training objective is to maximize the average log probability:

Sentence:	I	had	the	withdrawal	headaches	for	a	couple	of	days	.
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Semantic type:	people	na	na	na	event	na	na	na	na	unit	eos

Figure 2. Annotation example

Table 2. Statistics of annotated data

Statistics	Training Set	Test Set
No. of sentences	4000	1000
No. of words	84 830	21 036
No. of unique words	8108	3405
No. of body part mentions	1002	258
No. of chem mentions	2181	538
No. of device mentions	1545	406
No. of event mentions	1784	545
No. of flavor mentions	176	13
No. of num mentions	1353	389
No. of people mentions	7607	1888
No. of time mentions	1019	221
No. of unit mentions	227	201

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t).$$

The resulting model obtains an array of semantic vectors, also known as word embeddings, containing predicted neighbor words of each word in the corpus. In our experiment, we used a 50-dimensional word embedding to represent a unique word. A 50-dimensional word embedding is composed of 50 words that are most likely to appear around the word of interest. Word embedding enables us to represent medical entities in social media text when their semantic information is similar to that of standard medical terms. To avoid rare words appearing in word embedding and negatively affecting the model performance, we pruned the vocabulary by replacing the less frequent words with a unified symbol, "UNK" (short for "unknown token"). We kept the top 5000 frequent words in their original form and replaced the remaining words with UNK. After the training, we generated a 50-dimensional embedding model.

Bi-LSTM RNN model

Our research objective, to identify the entity types related to e-cigarette adverse events, can be considered as an NER task. In online forums, many user-generated posts contain long sentences. Besides, the semantic meaning of a word can be influenced by the words before and after it. Motivated by this intuition, we designed a language model that can handle long sentences and process sentences both forward and backward, thus capturing the previous and future word information at the same time. To this end, we developed a Bi-LSTM model to extract medical entities from the online forum text.

Given an input vector $(x_1,\ldots\ldots,x_T)$, an RNN computes the output $(y_1,\ldots\ldots,y_T)$ by iterating the following equations:

$$h^{(t)} = \mathrm{sigm}\left(Ux^{(t)} + \mathrm{W}h^{(t-1)}\right);$$
$$o^{(t)} = Vh^{(t)}.$$

sigm is the sigmoid function; U, W, and V are the weight vectors. At each time step, the RNN takes the last hidden state $h^{(t-1)}$ and the current input $x^{(t)}$ to compute the current hidden state $h^{(t)}$, and it uses the current hidden state $h^{(t)}$ to compute the current output $o^{(t)}$. The current hidden state $h^{(t)}$ is further passed to the next iteration to calculate the next hidden state $h^{(t+1)}$.

LSTM is a unique RNN architecture. Each LSTM unit contains an input gate $i^{(t)}$, a forget gate $f^{(t)}$, an output gate $o^{(t)}$, a memory cell $c^{(t)}$, and a hidden state $h^{(t)}$. The LSTM unit computes the output by iterating the following equations:

$$i^{(t)} = \mathrm{sigm}\left(W_i x^{(t)} + U_i h^{(t-1)} + b_i\right);$$
$$f^{(t)} = \mathrm{sigm}\left(W_f x^{(t)} + U_f h^{(t-1)} + b_f\right);$$
$$o^{(t)} = \mathrm{sigm}\left(W_o x^{(t)} + U_o h^{(t-1)} + b_o\right);$$
$$u^{(t)} = \tanh\left(W_u x^{(t)} + U_u h^{(t-1)} + b_u\right);$$
$$c^{(t)} = i^{(t)} \odot u^{(t)} + f^{(t)} \odot c^{(t-1)};$$
$$h^{(t)} = o^{(t)} \odot \tanh(c^{(t)}).$$

$x^{(t)}$ is the input at time step t. $\odot$ denotes element-wise multiplication. W, U, and b are the weight vectors of the gate parameters. The forget gate controls the extent to which the previous memory cell is forgotten, the input gate controls how much each unit is updated, and the output gate controls the exposure of the internal memory state.[45]

A Bi-LSTM[46] consists of 2 LSTMs that run in parallel: 1 on the input sequence and the other on the reverse of the input sequence. At each time step, the hidden state of the Bi-LSTM is the concatenation of the forward and backward hidden states. This setup allows the hidden state to capture both past and future information.[45] To reduce computational complexity, we trained a 50-dimensional word embedding model, meaning each word was converted to a 50-dimensional semantic vector. Then the word sequence was represented as an embedding sequence, which was passed to the Bi-LSTM layer. Instead of using a large hidden layer size, we used 150 neurons in the Bi-LSTM layer to avoid overfitting. This hidden layer size setup has also been successfully tested in other studies.[47] The outputs of the Bi-LSTM layers were then processed to a Softmax classifier, which predicts the entity type of each word in the input sentence. A graphic illustration of our Bi-LSTM is shown in Figure 3.

The Bi-LSTM model was trained on the annotated 4000 sentences, with 1000 sentences as the validation set (for cross-validation). Another 1000 annotated sentences were used as the test set. The model can predict the entity type of each word in the sentence automatically. The entity types are shown in Table 1.

Baseline models

Statistical learning method: conditional random fields

CRFs, a class of undirected statistical graphical model, have been widely adopted as the state-of-the-art NER method.[48–50] We used CRFsuite[51] as the implementation for our CRF baseline because it is fast. It can generate features automatically given the training text and provides a simple interface for training the input features.[48,51] The CRF classifier can predict the entity types given new sentences.

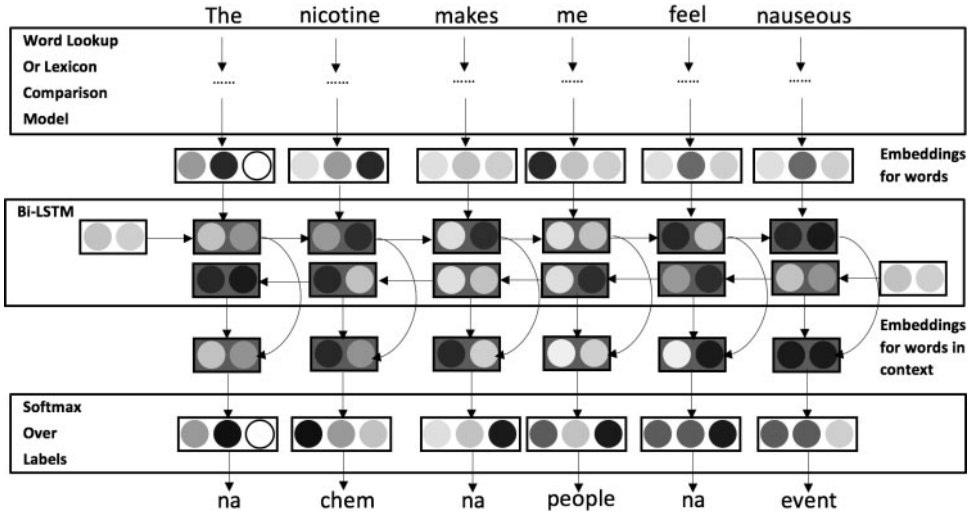

Figure 3. The Bi-LSTM RNN architecture

Table 3. UMLS entity types

Entity Type	UMLS Entity Type
bodypart	bdsy (Body System), blor (Body Location or Region), bpoc (Body Part, Organ, or Organ Component)
chem	chem (Chemical), chvf (Chemical Viewed Functionally), chvs (Chemical Viewed Structurally), clnd (Clinical Drug), elii (Element, Ion, or Isotope), enzy (Enzyme), hops (Hazardous or Poisonous Substance), inch (Inorganic Chemical), orch (Organic Chemical), phsu (Pharmacologic Substance)
device	drdd (Drug Delivery Device), medd (Medical Device)
event	acab (Acquired Abnormality), dsyn (Disease or Syndrome), inpo (Injury or Poisoning), mobd (Mental or Behavioral Dysfunction), patf (Pathologic Function), sosy (Sign or Symptom)
flavor	No match
num	qnco (Quantitative Concept)
people	humn (Human), famg (Family Group)
time	tmco (Temporal Concept)
unit	No match

Lexicon-based methods

MetaMap. MetaMap[52] is a Java application programming interface that accesses the UMLS, a medical dictionary maintained by the National Library of Medicine. Many medical studies have used MetaMap to find biomedical concepts from text.[30,53,54] We used MetaMap to identify the entity types of the words given the input sentences in the test set. We selected 21 entity types in the MetaMap entity type option, which are shown in Table 3.

MetaMap + consumer health vocabulary. CHV[55] complements the existing UMLS framework and helps to interpret consumer health vocabularies. CHV covers all entity types in the UMLS and enables the translation of consumer language to professional technical terms. We utilized CHV to map medical terms in informal and non-technical language to standard medical terms and then MetaMap to identify their entity types.

RESULTS

Evaluation

To evaluate the performance of our proposed model, we adopted precision, recall, and F-measure metrics. We compared our model with 3 strong baseline models: lexicon-based named entity recognition with MetaMap and CHV, and the state-of-the-art statistical learning model CRF. The annotated data was divided into 2 parts: 80% for training and 20% for testing. The precision (P), recall (R), and F-measure (F) for all entity types are shown in Table 4. The bold numbers are the best performance for each entity type.

Our model achieved much higher recall and F-measure than the 3 baseline models for all entity types, mainly due to the capability of recognizing adverse event entities with variations and rare adverse events in social media. Combining MetaMap and CHV achieved worse performance than using MetaMap alone, mainly because of the semantic drift that CHV caused. Many common words were extracted as medical entities of interest based on CHV when they should not have been. For instance, CHV converted "us" to "the United States," "an" to "autonomic nervous system," and "me" to "chronic fatigue syndrome." This negatively affected the accuracy of the system.

E-cigarette–related concepts

We applied our Bi-LSTM model to the entire corpus. Together with the expert-annotated dataset, we identified 1591 unique adverse events and 9930 unique e-cigarette components (ie, chemicals,

Table 4. Experiment results

Entity Type	MetaMap			MetaMap + CHV			CRF			Bi-LSTM RNN		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
All	42.5	21.7	28.7	26.9	18.6	22.0	**95.9**	75.3	84.4	94.1	**91.8**	**92.9**
Body part	75.4	55.3	63.8	35.7	22.9	27.9	**95.2**	67.0	78.6	93.9	**89.8**	**91.8**
Chemical	67.8	35.2	46.3	30.0	33.6	31.7	**98.9**	65.9	79.1	82.1	**91.5**	**86.5**
Device	4.9	0.7	1.2	4.8	0.7	1.2	**99.1**	79.0	87.9	94.4	**91.9**	**93.1**
Event	65.6	45.6	53.8	31.2	42.5	36.0	**99.5**	38.6	55.6	91.9	**77.3**	**84.0**
Flavor	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**	6.7	12.6	84.6	**73.3**	**78.6**
Number	59.3	100.0	74.5	18.5	37.8	24.8	78.5	94.6	85.8	**96.5**	**98.4**	**97.4**
People	57.9	2.3	4.4	47.6	1.1	2.2	**99.8**	91.1	95.3	98.6	**96.7**	**97.6**
Time	29.0	83.8	43.1	29.4	83.3	43.5	**99.4**	72.5	83.8	95.7	**89.1**	**92.3**
Unit	0.0	0.0	0.0	0.0	0.0	0.0	70.3	31.8	43.8	**88.2**	**82.5**	**85.3**

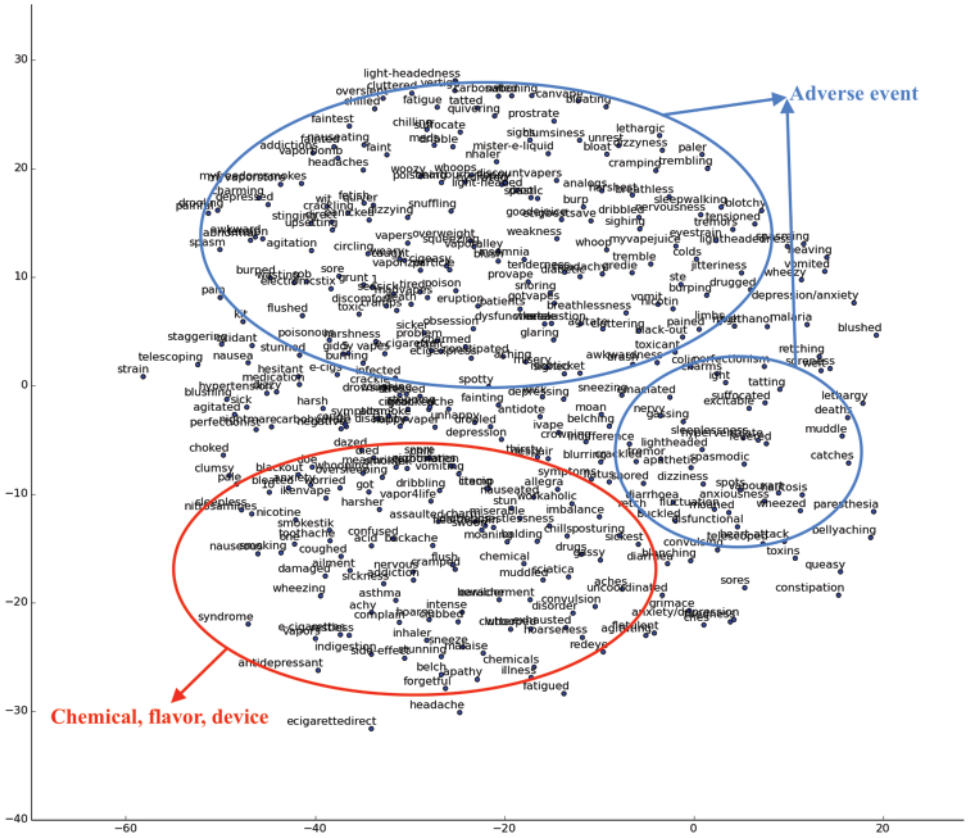

Figure 4. Word embedding visualization for e-cigarette related entities

flavors, and devices) from the entire research corpus. As mentioned before, word embedding contains semantic information of words. Words with similar semantic meaning have similar vector representation. To demonstrate this feature of word embedding, we visualized the semantic similarity of words based on the embedding representations. We used the t-SNE technique,[56] which reduces the dimensions of the embedding from 50 to 2 while preserving the relevant distance among the vectors. Words related to e-cigarette are plotted in Figure 4.

In the word embedding visualization, semantically similar words are clustered together. For instance, most words in the red circle are about adverse events, words in the blue circle are basically about chemicals, and words in the green circle are related to flavors. This result indicates that word embedding represents semantically similar words with similar vectors, and this representation is invariant to different spellings of words.

There are 34 287 adverse event entities (1591 unique ones). The adverse event entities account for 8.49% of all the extracted entities. Cough, headache, allergy, asthma, sore throat, and migraine were very commonly reported among e-cigarette users. Allergy, eye-twitch, fatigue, and asthma, which can potentially lead to serious health outcomes, have not been noted by the FDA. The new reports of adverse e-cigarette events will be valuable to the FDA's product safety monitoring program. We also identified 59 597 chemical entities (6509 unique ones; ie, vg, pg, caffeine, nicotine), 2879 flavor entities (334 unique ones; ie, chocolate, cherry, banana, vanilla), and

36 548 device entities (3087 unique ones; ie, madvapes, cloupor, anjelvape77, vapemail). Chemical entities account for 14.79% of all the extracted entities, flavor entities account for 0.71%, and device entities account for 9.05%.

DISCUSSION

Findings

This research aimed to extract adverse events related to e-cigarette from social media content. We developed the Bi-LSTM model with word embedding as the input representation. Although our model had slightly lower precision than the CRF model, it achieved much higher recall, resulting in the best F-measure among 3 strong baseline models. Our proposed method addresses the issues of the existing entity recognition methods. Our evaluation results show that our model reaches a precision of 94.10%, a recall of 91.80%, and an F-measure of 92.94%. The recall is 16% higher than the state-of-the-art CRF method, and the F-measure is 8% higher than CRF. The high recall ensures that our model can detect most of the relevant adverse events from the corpus data. We detected e-cigarette–related entities such as adverse events, chemical compounds, flavors, and devices. Some adverse events that we identified have not been noted by the FDA yet, including allergy, eye-twitch, fatigue, and asthma. Since the FDA has just started to regulate e-cigarettes, the agency has received limited e-cigarette safety reports. Social media, however, has matured and accumulated a large volume of e-cigarette discussions and feedback. In this sense, this data source provides valuable insights that are unnoted by the FDA.

Implications

First, our proposed Bi-LSTM is very useful in extracting medical entities from user-generated content. Compared to other entity recognition methods, our method achieved much higher recall, meaning our model can identify medical entities that have typos, abbreviations, and other variations in social media content. This is because the Bi-LSTM with word embedding is able to capture and process the semantic meanings of words. Furthermore, our approach, which uses social media data, can assist e-cigarette postmarket surveillance and increase the understanding of users' experiences with e-cigarettes. Our method automatically identifies discussions about adverse events, chemical compounds, e-cigarette device parts, and brands with high accuracy. This information can help clinical practitioners see from the consumers' perspective and gain better knowledge about emerging issues in the e-cigarette market. These user experiences will complement clinical experiments and enrich the knowledge of e-cigarette safety issues. Moreover, our research can improve consumer awareness of harmful outcomes of e-cigarette use. While most campaigns promote e-cigarettes as a benign alternative to conventional cigarettes or other tobacco products, few new consumers are aware that e-cigarette use can cause adverse events. Our findings can also provide supplemental information for regulatory agencies to improve consumer awareness of harmful outcomes of e-cigarettes.

Limitations

First, we trained a Bi-LSTM model to reduce computational complexity. To obtain higher performance, we can train an LSTM network, though at the cost of longer training duration. Second, the word embedding model contained 50 dimensions. Higher-dimensional embedding models can be trained to capture more accurate semantic information. Third, we did not consider the relationships between adverse events and chemical compounds. Relation extraction can be applied to identify the adverse events related to a particular chemical compound. Fourth, social media surveillance alone is not enough for comprehensive e-cigarette safety regulation. Clinical experiments, together with complementary online surveillance, will improve clinical implications. For future research, more sophisticated models can be tested and further analysis such as relation extraction can be performed.

CONCLUSION

E-cigarettes have been proven to cause adverse effects. However, previous medical studies and e-cigarette safety monitoring systems failed to identify adverse e-cigarette events on a large scale. This study aimed to extract e-cigarette–related information from a large volume of social media data. We developed a high-performance information extraction framework for e-cigarette social media safety surveillance using a deep neural network method. Although the CRF baseline model had slightly better precision, our Bi-LSTM RNN model achieved much higher recall, resulting in a higher F-measure than strong statistical learning and lexicon-based baselines. To the best of our knowledge, we are among the first to develop a deep neural network–based approach for understanding medical information in social media, and this is also the first study to examine e-cigarette safety issues on a large scale. By incorporating the LSTM unit and the bidirectional architecture, our proposed Bi-LSTM model can accurately extract relevant medical entities in social media data. This framework can be generalized to solve other problems, such as adverse drug event detection and drug-drug interactions. Based on the extracted information, we identified adverse events unnoted by the FDA and prior studies, which further demonstrates the value of user-generated social media content for e-cigarette safety surveillance. Our research supports tobacco product regulatory policy makers by providing new evidence of harmful e-cigarette effects, such as allergy, eye-twitch, fatigue, and asthma. We also contribute to health informatics research by designing a novel computational method for named entity recognition. Future research can focus on finding the best configuration of model parameters, such as the number of hidden layers and word embedding dimensions.

FUNDING

This work is supported by the US National Institutes of Health (grant no. 1R01DA037378-01) and National Science Foundation (grant nos. IIS-1553109 and IIS-1552860).

CONTRIBUTORS

JX designed the study, collected the data, built the model, and ran the experiment. XL helped design the study. This study was conducted under the supervision of DDZ. DDZ provided substantial contributions to study design and research funding.

COMPETING INTERESTS

None.

IRB STATEMENT

This study does not require Institutional Review Board oversight by the determination of the University of Arizona and the University of Utah.

REFERENCES

1. Dutra LM, Glantz SA. Electronic cigarettes and conventional cigarette use among US adolescents: a cross-sectional study. *JAMA Pediatrics*. 2014;168(7):610–17.
2. Schoenborn CA, Gindi RM. Electronic cigarette use among adults: United States, 2014. *NCHS Data Brief*. 2015;217:1–8.
3. FDA. *Secondary* 2015. http://www.fda.gov/NewsEvents/PublicHealthFocus/ucm172906.htm. Accessed April 2, 2017.
4. Research B. *Electronic Cigarette & E Vapor (Vaporizer) Market Research Reports. Secondary Electronic Cigarette & E Vapor (Vaporizer) Market Research Reports* 2016. http://bisresearch.com/electronic-cigarette-market-size-forecast.html. Accessed April 2, 2017.
5. Callahan-Lyon P. Electronic cigarettes: human health effects. *Tobacco Control*. 2014;23(Suppl 2):ii36–ii40.
6. Chen I-L. FDA summary of adverse events on electronic cigarettes. *Nicotine Tobacco Res*. 2013;15(2):615–16.
7. FDA. FDA takes significant steps to protect Americans from dangers of tobacco through new regulation. *Secondary FDA takes significant steps to protect Americans from dangers of tobacco through new regulation* 2016. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm499234.htm. Accessed April 2, 2017.
8. Huerta TR, Walker DM, Mullen D, Johnson TJ, Ford EW. Trends in E-Cigarette Awareness and Perceived Harmfulness in the US. *Am J Prevent Med*. 2016;52(3):339–46.
9. Palazzolo DL. Electronic cigarettes and vaping: a new challenge in clinical medicine and public health. A literature review. *Front Public Health*. 2013;1:56.
10. Westenberger B. *Evaluation of e-cigarettes*. Food and Drug Administration: St Louis, MO; 2009:1–8.
11. Polosa R, Caponnetto P, Morjaria JB, Papale G, Campagna D, Russo C. Effect of an electronic nicotine delivery device (e-cigarette) on smoking reduction and cessation: a prospective 6-month pilot study. *BMC Public Health*. 2011;11(1):1.
12. Bullen C, McRobbie H, Thornley S, Glover M, Lin R, Laugesen M. Effect of an electronic nicotine delivery device (e cigarette) on desire to smoke and withdrawal, user preferences and nicotine delivery: randomised crossover trial. *Tobacco Control*. 2010;19(2):98–103.
13. Goniewicz ML, Knysak J, Gawron M, *et al*. Levels of selected carcinogens and toxicants in vapour from electronic cigarettes. *Tobacco Control*. 2014;23(2):133–39.
14. Flouris AD, Chorti MS, Poulianiti KP, *et al*. Acute impact of active and passive electronic cigarette smoking on serum cotinine and lung function. *Inhalation Toxicol*. 2013;25(2):91–101.
15. Tzatzarakis MN, Tsitoglou KI, Chorti MS, *et al*. Acute and short term impact of active and passive tobacco and electronic cigarette smoking on inflammatory markers. *Toxicol Lett*. 2013(221):S86.
16. Vansickel AR, Cobb CO, Weaver MF, Eissenberg TE. A clinical laboratory model for evaluating the acute effects of electronic "cigarettes": nicotine delivery profile and cardiovascular and subjective effects. *Cancer Epidemiol Biomarkers Prevent*. 2010;19(8):1945–53.
17. Vardavas CI, Anagnostopoulos N, Kougias M, Evangelopoulou V, Connolly GN, Behrakis PK. Short-term pulmonary effects of using an electronic cigarette: impact on respiratory flow resistance, impedance, and exhaled nitric oxide. *Chest J*. 2012;141(6):1400–06.
18. Ji Y, Ying H, Dews P, *et al*. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Trans Inf Technol Biomed*. 2011;15(3):428–37.
19. Farsalinos KE, Polosa R. Safety evaluation and risk assessment of electronic cigarettes as tobacco cigarette substitutes: a systematic review. *Therapeutic Adv Drug Safety*. 2014;5(2):67–86.
20. FDA. MedWatch Online Voluntary Reporting Form. *Secondary MedWatch Online Voluntary Reporting Form* 2016. https://www.accessdata.fda.gov/scripts/medwatch/index.cfm?action=reporting.home. Accessed April 2, 2017.
21. Derczynski L, Maynard D, Rizzo G, *et al*. Analysis of named entity recognition and linking for tweets. *Inform Process Manag*. 2015;51(2):32–49.
22. Liu X, Chen H. Identifying adverse drug events from patient social media: a case study for diabetes. *IEEE Intell Sys*. 2015;30(3):44–51.
23. Wang C, Zimmermann MT, Prodduturi N, Chute CG, Jiang G. Adverse drug event-based stratification of tumor mutations: a case study of breast cancer patients receiving aromatase inhibitors. *AMIA Annual Symposium Proceedings. American Medical Informatics Association*; 2014;2014:1160.
24. Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C. Focus on clinical care and patient safety: Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. *J Am Med Inform Assoc*. 2011;18(Suppl 1):i73.
25. Sarker A, Ginn R, Nikfarjam A, *et al*. Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform*. 2015;54:202–12.
26. Greene JA, Kesselheim AS. Pharmaceutical Marketing and the New Social Media. *New Engl J Med*. 2010;363(22):2087–89.
27. NLM. *Unified Medical Language System. Secondary Unified Medical Language System* 2009. https://www.nlm.nih.gov/research/umls/. Accessed April 2, 2017.
28. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol*. 2013;9(2):e1002854.
29. Friedman C. A broad-coverage natural language processing system. *Proceedings of the AMIA Symposium. American Medical Informatics Association*; 2000:270.
30. Osborne JD, Gyawali B, Solorio T. Evaluation of YTEX and MetaMap for Clinical Concept Recognition. arXiv preprint arXiv:1402.1668 2014.
31. Gupta S, MacLean DL, Heer J, Manning CD. Induced lexico-syntactic patterns improve information extraction from online medical forums. *J Am Med Inform Assoc*. 2014;21(5):902–09.
32. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. *J Am Med Inform Assoc*. 2014;21(5):808–14.
33. Li K, Ai W, Tang Z, *et al*. Hadoop recognition of biomedical named entity using conditional random fields. *IEEE Trans Parallel Distributed Sys*. 2015;26(11):3040–51.
34. Wei Q, Chen T, Xu R, He Y, Gui L. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database*. 2016;2016:baw140.
35. Benson E, Haghighi A, Barzilay R. Event discovery in social media feeds. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics*. 2011;1:389–98.
36. Jakob N, Gurevych I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*. 2010:1035–45.
37. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. *Neural Architectures for Named Entity Recognition*. arXiv preprint arXiv:1603.01360 2016.
38. Rubenstein H, Goodenough JB. Contextual correlates of synonymy. *Commun ACM*. 1965;8(10):627–33.
39. Dos Santos CN, Zadrozny B. *Learning Character-level Representations for Part-of-Speech Tagging*. ICML; 2014:1818–26.
40. Baldwin T, Kim Y-B, de Marneffe MC, Ritter A, Han B, Xu W. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP*;2015:126.
41. Sutskever I, Vinyals O, LeQV. Sequence to sequence learning with neural networks. *Adv Neural Inform Process Sys*. 2014:3104–12.

42. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. *ICML (3)*. 2013;28:1310–18.
43. Blackman NJM, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Stats Med*. 2000;19(5):723–41.
44. Milokov T, Sutskever I, Chen K, Corrando GS, Dean J. Distributed representations of words and phrases and their compositionality. *Adv Neural Inform Process Sys*. 2013:3111–19.
45. Tai KS, Socher R, Manning CD. *Improved Semantic Representations from Tree-structured Long Short-term Memory Networks*.arXiv preprint arXiv:1503.00075 2015.
46. Graves A, Jaitly N, Mohamed AR. Hybrid speech recognition with deep bidirectional LSTM. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE; 2013.
47. Ling W, Luís T, Marujo L, *et al*. *Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation*.arXiv preprint arXiv:1508.02096. 2015.
48. Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*. 2010:384–94.
49. Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*; 2011:1524–34.
50. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symp Biocomput*. 2008;13:652–63.
51. Okazaki N. *CRFsuite: a Fast Implementation of Conditional Random Fields (CRFs)*. 2007. http://www.chokkan.org/software/crfsuite/. Accessed April 2, 2017.
52. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium. American Medical Informatics Association*. 2001:17.
53. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229–36.
54. Hanauer DA, Saeed M, Zheng K, *et al*. Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: a feasibility analysis. *J Am Med Inform Assoc*. 2014;21(5):925–37.
55. Utah UO. *Collaborative Consumer Health Vocabulary Initiative. Secondary Collaborative Consumer Health Vocabulary Initiative* 2011. http://consumerhealthvocab.org/. Accessed April 2, 2017.
56. Maaten Lvd, Hinton G. Visualizing data using t-SNE. *J Machine Learning Res*. 2008;9(Nov):2579–605.