Social Media-based Overweight Prediction Using Deep Learning

Abstract

Overweight is epidemic in the United States and elsewhere in the world, causing major health concerns. Based on self-disclosure theory, i.e., people have the tendency to disclose information concerning their feelings, intentions, and acts (e.g., food consumption) online, we aim to leverage social media platforms to develop an unobtrusive approach to predicting overweight. However, traditional statistical and machine learning-based approaches either deliver unsatisfactory performance or demand a large number of features. In this paper, we present a novel social media-based overweight prediction approach based on deep learning as applied in the context of Natural Language Processing (NLP). The input to this approach is food-related Twitter posts. Our computational results show the effectiveness of our method, with remarkable improvement in terms of accuracy over a set of benchmark methods.

Keywords

Public health, overweight, food, language, Twitter, deep learning.

Introduction

Overweight is a global epidemic health problem. Overweight refers to that the value of Body Mass Index (BMI), a measure of weight relative to height (Eknoyan 2008), is equal to or higher than 25 kg/m². Overweight is also a precursor to obesity. Obesity refers to the case when BMI is equal to or higher than 30 kg/m². According to World Health Organization (WHO), 39% of men and 39% of women (nearly 2 billion adults) were overweight, and 11% of men and 15% of women obese (more than half a billion) in 2016. Overweight may lead to diseases such as heart diseases, stroke, gallstones, cancers, and especially type 2 diabetes mellitus (T2DM). Moreover, obesity has been regarded as a chronic progressive disease. Overweight and related issues can have a serious impact on human health.

---

Social Media-based Overweight Prediction Using Deep Learning

It has also been shown that overweight is positively correlated with the residential income level of countries (Jolliffe 2011), especially in industrialized nations. In the United States, overweight and obesity continue to be severe health issues. Approximately 78.6 million U.S. adults are obese, and about twice are overweight. Also, about 12.7 million children and adolescents, aged between two to nineteen, suffer from obesity. Figure 1 above shows the perceived severity of the problem in the U.S.

In addition to causing health problems, overweight also results in a significant amount of health expenditure. For example, the cost for treating T2DM, a chronic and debilitating disease, which highly correlates with overweight, mounted up to $245 billion in 2012. Any reduction in overweight rates would lead to massive healthcare savings.

Online social media provide predictive power and numerous opportunities for studying such serious and costly public health-related issues from both theoretical and practical perspectives. Theoretically, based on self-disclosure theory, information related to food online can be leveraged to predict overweight. Practically, a variety of previous studies have successfully used information extracted from social media to monitor diseases such as flu (Ginsberg et al. 2009), diabetes (Greene et al. 2011) and cancer (Sugawara et al. 2012). In particular, Twitter has been utilized as a popular source in the previous studies on public health because of its vast amount of information and high accessibility.

However, traditional methods, such as statistical approaches and machine learning approaches (e.g., Support Vector Machine (SVM)), are often not adequate in processing social media big data and delivering satisfactory predictions. In many cases, their performance as typically measured by accuracy is low (Fried et al. 2014). In some cases where such approaches perform well, a large number of features will

---

Figure 1. Prevalence of Obesity in the United States in 2016

8 "More than 29 million Americans have diabetes; 1 in 4 doesn't know," Internet: http://www.cdc.gov/media/releases/2014/p0610-diabetes-report.html (last accessed 5 February 2018)
be required (De Choudhury et al. 2016; Fried et al. 2014). These challenges motivate this reported research. We aim at answering the following key research question:

**How can we better predict overweight based on online social media content?**

We address this question by developing a deep learning-based approach to predicting overweight and the factors correlated with overweight using Twitter posts.

We chose Twitter as our study context for three reasons. First, Twitter is widely used across various ethnic, gender, age, and social-economic groups (Preotiuc-Pietro et al. 2017). Second, users prefer to write short paragraphs to talk about daily activities, including diet (Paul and Dredze 2011). Third, a huge number of records of eating-related descriptions and discussions are traceable on Twitter, which provides researchers a highly accessible and potentially cost-effective means of acquiring public health-related data.

We employed the deep learning framework because it has been proven to be effective in greatly improving predictive accuracy in a wide range of contexts by learning hidden representations (LeCun et al. 2015). For example, it can help us learn latent features or variables that are used to capture dependencies of concern. Deep learning has already been applied with great results in many streams of Natural Language Processing (NLP) and text mining research, such as document classification (Lai et al. 2015) and document summarization (Cao et al. 2015). Our findings support the effectiveness of the approach, with at least 8% absolute improvement of accuracy than a set of benchmark methods. These results reveal that our method is capable of effectively monitoring overweight issues from Twitter.

Specifically, our contributions to the IS field are:

- A research framework that has the potential to be valuable not only in overweight prediction but also other health-related predictions;
- A novel social media-based overweight prediction approach based on deep learning as applied in the context of NLP; and
- A study that empowers the applications of design science.

The remainder of the paper is organized as follows. In Section II, we discuss the theoretical foundation (i.e., self-disclosure theory) of our work. In Section III, we review related work. We then propose a research framework based on deep learning for overweight prediction. We also provide details about data collection, preparation, lexical semantic extension, classification, evaluation and our results in Section IV. The ensuing Section V discusses findings and concludes with a summary of our research contributions and potential future directions.

**Self-Disclosure Theory**

Self-disclosure is defined as the behavior that individuals voluntarily reveal their information to others, including daily activities and feelings (Qian and Scott 2007). The benefits for self-disclosure include the formation of intimate associations, social recognition, and social anxiety management (Green et al. 2016).

Self-disclosure exists not only in face-to-face communication (e.g., offline) but also in computer-mediated communication (e.g., online) (Mioso 2015). Some studies have shown that the willingness to disclose information is higher in online context than in offline context (Joinson 2001; Bargh et al. 2002).

Online self-disclosure information facilitates social studies to be conducted in an unobtrusive way through social media. Information about food, drink, and medical status is widely available online because of self-disclosure (Lin et al. 2015; Ma et al. 2017; Wang et al. 2016).

Based on the self-disclosure theory, we assume that information related to food disclosed by oneself online can reveal one's food preference, and in turn, be leveraged to predict overweight. The food-related information does not, of course, necessarily indicate overweight due to some inherent biases in Twitter data caused by self-reported nature (e.g., people may talk more about food because of interests rather than real food consumption) and the correlation between food-related information and overweight is only meaningful across large populations. Therefore, we can use online social media information to conduct data analytics for overweight prediction at state-level. Moreover, we include food-related features (not just food words) to provide us a more comprehensive understanding of the situation.
Related Work

This paper is closely related to three important streams of literature. The first stream expounds the significance and popularity of Twitter for monitoring public health. The second stream is related to approaches applied to public health monitoring. The third stream is concerned about deep learning from a methodological point of view.

Twitter has been utilized as a popular platform and data source for public health monitoring (George and Dellasega 2011; Gupta and Kohli 2016; Paul et al. 2011; Fernandez-Luque et al. 2010; Hingle et al. 2013; McKendrick et al. 2012; Yom-Tov et al. 2014). It enables researchers to tap into the wealth and richness of user-generated contents, regardless of temporal or geographical boundaries. The informal and colloquial nature of Twitter posts, along with its ease of data access, make it chosen by many works. Yom-Tov et al. (2014) extracted Twitter postings about mass gatherings (e.g., festivals and religious events) to detect disease outbreaks. Menon (2006) conducted the content analysis on tweets, which were crawled based on keywords like “H1N1” and “swine flu” to monitor flu. Schwartz et al. (2013) analyzed tweets to understand life satisfaction of users across different demographic and social statuses.

The most common approaches for public health monitoring include statistical and machine learning-based approaches (Abbasi and Adjeroh 2014; Khoury and Ioannidis 2014; Brady et al. 2017; Ginsberg et al. 2009; Hingle et al. 2013). For example, Yom-Tov et al. (2014) calculated the probability of words occurrence before and after the date of an event and utilized a statistical test to determine significance. Ginsberg et al. (2009) detected influenza epidemics by figuring out the change in large numbers of Google search queries. Hingle et al. (2013) leveraged Twitter together with analytical methods to capture food-related activities based on a very small dataset. Others relied on machine learning approaches such as topic modeling (Paul and Dredze 2011; Schwartz et al. 2013) and SVM (Fried et al. 2014). The study most relevant to ours is Fried et al. (2014). They collected a large corpus of over three million food-related tweets and thus predicted many latent population characteristics, including overweight. However, these existing approaches either lack good performance, or a large number of feature sets are required. Few works utilized deep learning in monitoring health issues.

Deep learning has recently emerged as a powerful NLP and text mining technique (Day and Lin 2017; Miotto et al. 2017; Pham et al. 2017; Ravi et al. 2017). Many studies demonstrated the improved prediction accuracies through the application of deep learning. Dong et al. (2014) developed a modified Cursive Neural Network (CNN) framework for modeling semantic composition and resulted in 3% improvement in sentence-level polarity classification. Similarly, Cao et al. (2015) presented a Ranking Recursive Neural Networks (R2N2) for document summarization and finally achieved higher scores than other state-of-the-art machine learning approaches. Lai et al. (2015) applied Recurrent Convolutional Neural Network (RCNN) to classify text, and it performed the best on three of the datasets. To the best of our knowledge, even though deep learning has been applied to many fundamental NLP or text mining research, it has not yet been employed in the identification of the overweight problem.

Methodology

Our research framework consists of five sequential modules as shown in Figure 2.
These five modules are data collection, data preprocessing, lexical semantics, classification, and evaluation. In Module I, data collection, raw data were collected based on food-related words as keywords and text extracted by regular expressions. In Module II, textual information was tokenized, stemmed, POS tagged and dependency extracted by means of NLP tools. Module III aims to extend lexical semantics based on Word2Vec. Module IV is mainly concerned with classification and its implementation. Finally, training and testing data were prepared, and the performance evaluated by 10-fold cross validation in Module V. More details about these modules are provided below.

**Data Collection**

We utilized the Twitter streaming API to collect tweets that contain food-related words. The food-related words were created based on the previous literature (Fried et al. 2014). We categorized them into four types: food and drink (Type I), context (including operations, tools) for food (Type II), people who are interested in food and their features (Type III), and characteristics of food (Type IV). We created 809 food-related words. Examples of some typical words are shown in Table 1. Finally, we collected 7.40 GB data from Twitter until May 1st, 2015, containing about 30 million sentences.

<table>
<thead>
<tr>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
<th>Type IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>blackberry</td>
<td>cook</td>
<td>appetizer</td>
<td>hot</td>
</tr>
<tr>
<td>beef</td>
<td>bake</td>
<td>chef</td>
<td>frozen</td>
</tr>
<tr>
<td>cabbage</td>
<td>boil</td>
<td>hungry</td>
<td>fried</td>
</tr>
<tr>
<td>chocolate</td>
<td>chopsticks</td>
<td>eat</td>
<td>greased</td>
</tr>
</tbody>
</table>

**Table 1. Examples of Food Related Words**
Data Preprocessing

In this step, we preprocessed data for the next analysis. After collecting all relevant tweets, we first implemented regular expressions to remove URL, HTTP, and other noisy information in the collected data. Then we regarded each tweet as a single unit and utilized Stanford NLP\(^9\) to deal with the standard processing of text, including tokenization, stemming, POS Tagging, and dependency tree generation. Finally, XML files with the lemma, POS and dependency tree information were automatically generated by the program.

Lexical Semantics

In the following analysis, we treated food-related words as base lexicon and employed Word2Vec\(^{10}\) to extend the lexicons. Word2Vec was developed by Mikolov et al. (2013), a typical model for lexical semantics and a pre-processing step for deep learning (Majumder et al. 2017). It uses raw text to learn semantic similarities between words and then create a matrix, in which each line of a word is a vector representation of other words with different values. We utilized Word2Vec to generate the similarity between two words based on the learned matrix. For a particular word, all the semantically related words were ranked from the highest similarity to the lowest. We then set a parameter to determine top N related words for that word, which is also known as Nearest-N words in Word2Vec. In our case, even though our dataset is very large, it is still sparse as for the appearance of many words. To resolve this issue, we utilized a trained matrix, which was based on the gigaword corpus\(^{11}\). This matrix uses the skip-gram variant of Word2Vec and produces vectors with 200 dimensions. Examples of the similarity between two words are illustrated in Table 2.

<table>
<thead>
<tr>
<th>Word1</th>
<th>Word2</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>brackish</td>
<td>bamboo</td>
<td>0.407</td>
</tr>
<tr>
<td>drinkable</td>
<td>ate</td>
<td>0.272</td>
</tr>
<tr>
<td>porcini</td>
<td>honey</td>
<td>0.762</td>
</tr>
</tbody>
</table>

Table 2. Examples of Lexical Similarity

Classification

After generating the word matrix, we built our classification model. There are 50 states and Washington D.C. in the United States. We treated Washington D.C. as a state. Our classification task was to label whether the population of a state is in general “overweight” or not based on the overweight rate of the national median. We note that although this reported work is on a state-level, the lessons learned have the potential to be used for individual-level predictions through transfer learning, which takes advantage of the “knowledge” learned from the state-level models and knowledge.

In our study, food-related words and extended words by Word2Vec were regarded as features. Then log normalization was used for features, which were stored as the matrix in CSV files. Deep Belief Network (DBN) Classifier (Hinton et al. 2006), which is able to handle numeric features, was utilized to classify whether a state was overweight or not.

---

10 “Word2Vec,” Internet: http://deeplearning4j.org/word2vec (last accessed 5 February 2018)
Evaluation

The performance of our model was assessed through 10-fold cross validation. Five or six of the states were regarded as testing data, and other states as training data. The average accuracy measures were then calculated. We tested the model starting from only containing food-related words, to extending the lexicons by each word’s Nearest-N words. The size of the neighborhood N was set as 10, 20, 30, 40, 50, 60, 70, and 80. Our evaluation results are shown in Figure 3. We can see the improvement of the accuracy with increasing lexical semantic extension with the help of Word2Vec, reaching 0.78 when N is 50.

Figure 3. Accuracy of Our Approach for N=10, 20, 30, 40, 50, 60, 70, 80

Main comparative results are shown in Table 3. In order to better interpret the performance of our approach, we compared our results with the state-of-the-art baselines including Logistic Regression (LR), SVM, and Decision Tree (DT). In addition, even though Word2Vec is usually used as the pre-processing step for deep learning (Majumder et al. 2017), we pair Word2Vec with the state-of-the-art baselines to make a fair comparison. Our approach outperforms these baselines.

<table>
<thead>
<tr>
<th>Model</th>
<th>Overweight Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food + LR</td>
<td>0.43</td>
</tr>
<tr>
<td>Food + Word2Vec + LR</td>
<td>0.61</td>
</tr>
<tr>
<td>Food + SVM</td>
<td>0.69</td>
</tr>
<tr>
<td>Food + Word2Vec + SVM</td>
<td>0.69</td>
</tr>
<tr>
<td>Food + DT</td>
<td>0.59</td>
</tr>
<tr>
<td>Food + Word2Vec + DT</td>
<td>0.57</td>
</tr>
<tr>
<td>Food + DBN</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Our Approach (Food + Word2Vec + DBN)</strong></td>
<td><strong>0.78</strong></td>
</tr>
</tbody>
</table>

Table 3. Comparison of Our Approach with Benchmark Methods
Our major finding from this study is that a good classifier and a viable feature set are both important. From Table 3, we can see that deep learning classifier performs best even without Word2Vec. With the help of Word2Vec, our approach can achieve at least 8% absolute improvement in terms of accuracy over a set of benchmark baselines, demonstrating the power of the proposed approach.

Word2Vec can be utilized to improve feature extraction and help gain better insights into the problem through learning the semantic meaning from a large amount of data, especially useful for deep learning algorithms. When Word2Vec is applied for LR and DBN, the results show 18% and 8% absolute improvement in terms of accuracy, respectively (Table 3). Word2Vec can help us extend words of people’s activities involved in cooking (e.g., “preheat”, “panfried”, “refried”, “diced”). It can also facilitate the identification of words associated with food (e.g., “grams”, “unsalted”, and “granulated”). Moreover, words like “mom”, “wife”, “friend” indicate that we’d like to have dinner with family or friends rather than alone. Words like “baker”, “foodie” and “artist” refer to people who are enthusiastic about eating or cooking. In sum, most of the words make sense in the extension process.

Discussion and Conclusion

Our research has important implications for both practitioners and academia. We proposed a novel research framework for overweight prediction. The framework advocates treating food-related words both as keywords for data collection and base lexicons for the further semantic extension. The framework also demonstrates how standard NLP and deep learning techniques can be integrated to solve a practical problem. Practitioners can implement our framework to better identify overweight issues. The performance of the proposed computational approach is encouraging as well. Application of this type of approach in other settings is worth pursuing.

We can improve this study from several aspects. First, in this work, we only consider food-related words. Actually, exercise-related words are also closely related to overweight. In our future study, such words will be included. Second, social cues including emotions, opinions, style, and topics can be considered to predict overweight. Also, we will apply the developed approach to individuals to infer overweight characteristics with the help of additional specific information (e.g., location, age, sex, and income).

REFERENCES


Learning for Health Informatics.” IEEE Journal of Biomedical and Health Informatics 21(1), pp. 4–21.