Deep Learning Approach for Overweight Prediction on Twitter

Luwen Huangfu

May 02, 2016

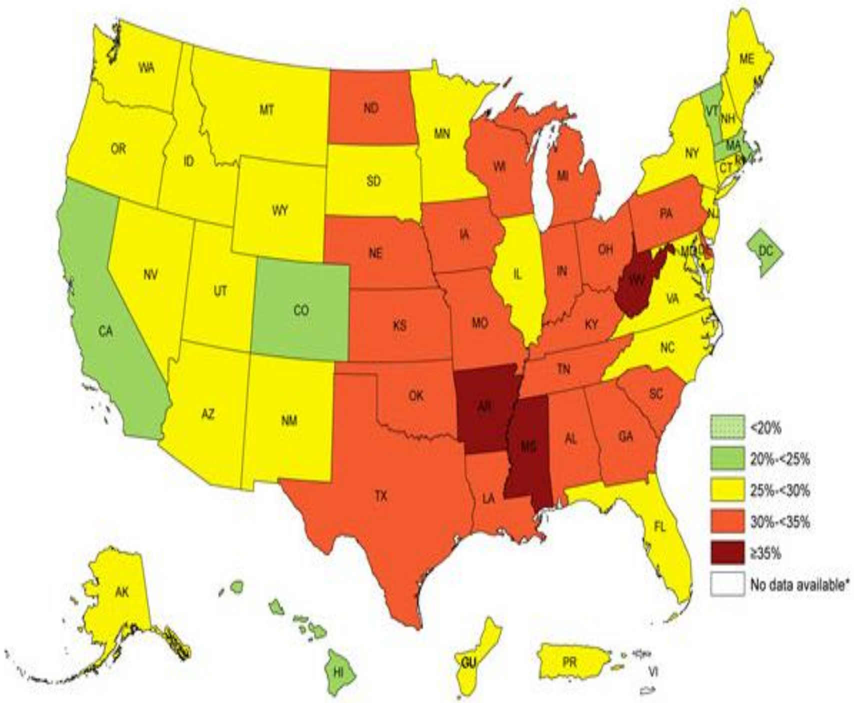
THE UNIVERSITY OF ARIZONA

Overview

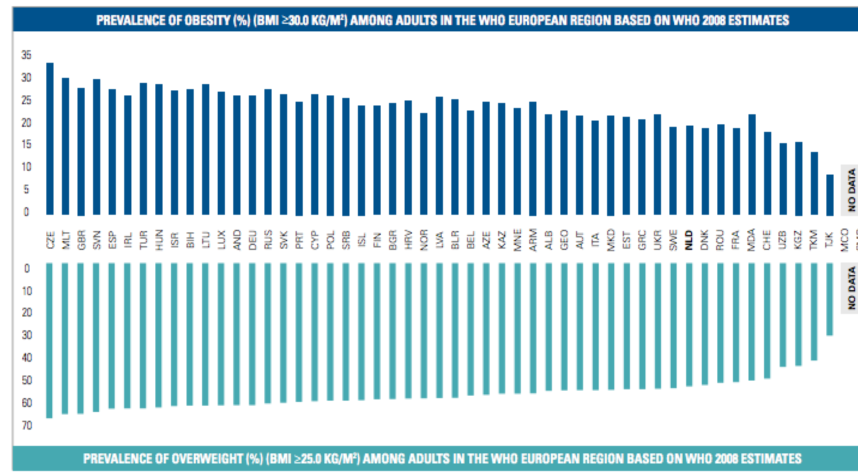
- Introduction
- Motivation
- Literature Review
- Research Question
- Research Design
- Research Experiment
- Findings and Discussions
- Conclusions and Future Directions

Introduction

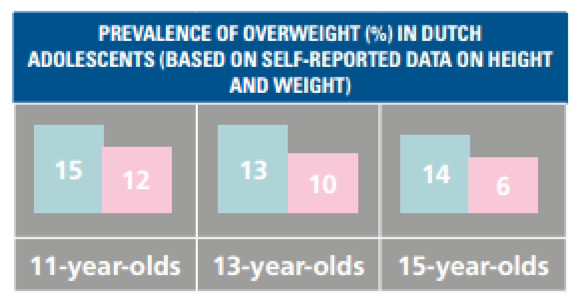
- Overweight is epidemic in the United States and elsewhere in the world



"Prevalence of self-reported obesity among U.S. adults by state and territory," Internet: http://www.cdc.gov/obesity/data/prevalence-maps.html, 2014



Notes. The country codes refer to the ISO 3166-1 Alpha-3 country codes. Data ranking for obesity is intentionally the same as for the overweight data. BMI: body mass index.
Source: WHO Global Health Observatory Data Repository *(1)*.



"Nutrition, physical activity and obesity: Netherlands," Internet: http://www.euro.who.int/__data/assets/pdf_file/0018/243315/Netherlands-WHO-Country-Profile.pdf?ua=1, 2013

Introduction

- Overweight is serious and costly nowadays
  - Cosmetic problem
  - Raise risks for other health problems
  - T2DM mounted up to $245 billion in 2012
- Overweight is closely related with diet and physical activities (Neel, 1999)
  - Energy intake > energy expenditure
  - Humans tend to store energy in case of famine
  - High-calorie food can lead to overweight

Motivation

- We are motivated to study overweight based on language of food on Twitter using deep learning approach
- Why study language of food on Twitter?
  - Across ethnic, gender, age, and social-economic groups
  - Short paragraphs are preferred to talk about daily activities
  - Permanent records of eating related behaviors
- Why deep learning approach?
  - Deep representations
    - e.g., learning intermediate concepts, features or latent variables that are useful to capture dependencies that we care about
  - Powerful in many research domains
    - e.g., image processing, video analysis
  - Emerge in NLP and text mining

Literature Review

- Twitter has been utilized as a popular source for public health monitoring
  - Track diseases (Yom-Tov et al., 2014; Chew et al., 2010)
  - Detect life satisfaction (Schwartz et al., 2013)
  - Identify overweight (Fried et al., 2014)
- Approaches for public health monitoring
  - Statistical Approach (Yom-Tov et al., 2014; Ginsberg et al., 2009)
  - Machine Learning Approach (Paul & Dredze, 2011; Schwartz et al., 2013; Fried et al., 2014)
- However, accuracy is not good or feature sets are really large
- Few work utilizes deep learning approach in monitoring health issues

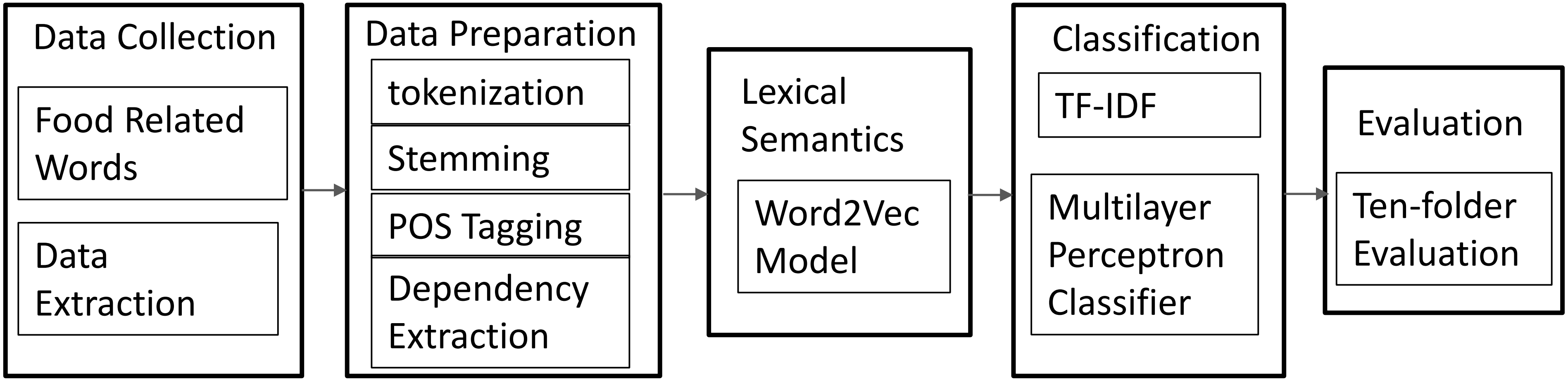
Literature Review

- Deep learning emerges as a powerful technique in NLP and text mining
  - Sentiment analysis (Dong et al., 2014)
  - Document summarization (Cao et al., 2015)
  - Text classification (Lai et al., 2015)
- Inspired by its power, we attempt to apply it to overweight identification, which has not been done by others

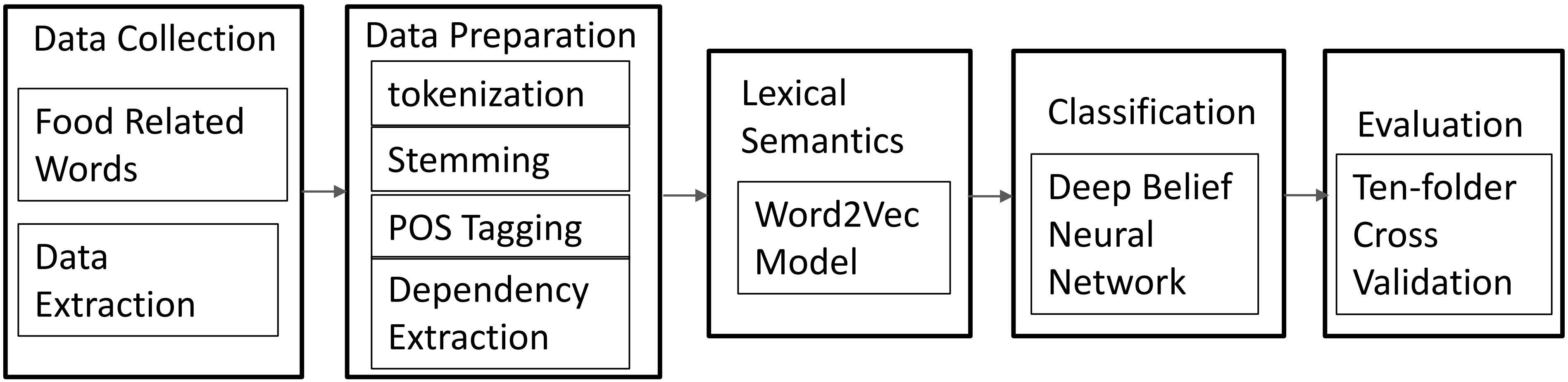
Research Question

- *Can we monitor overweight issue from Twitter by using deep learning approach?*

Research Design



Research Design



Data Collection

- Twitter Stream API
- Food related words (801 words)
- 7.40GB till May 1st, 2015
- 30,000,000+ sentences

Food Related Words Examples				
apple	ate	bacon	banana	barbecue
beef	biscuit	blackberry	blueberry	bread
cabbage	cake	carrot	cereal	cheese
chocolate	chopsticks	cocoa	coconut	coffee
		……		

Data Preparation

- Text extraction
  - Removal of URL, HTTP and other noisy information
- Text processing
  - http://nlp.stanford.edu/software/
  - Tokenization
  - Stemming
  - POS Tagging
  - Dependency Tree Generation

Lexical Semantics

- Food related words as base lexical
- Word2Vec for lexical extension
  - http://deeplearning4j.org/word2vec
  - Similarity calculation
  - Extend base lexical by nearest N-words

word1	word2	similarity
wholegrain	banana	0.43262019753456116
diningroom	appetizer	0.2730669677257538
kitchenette	honey	0.1797347962856292

Lexical Semantics

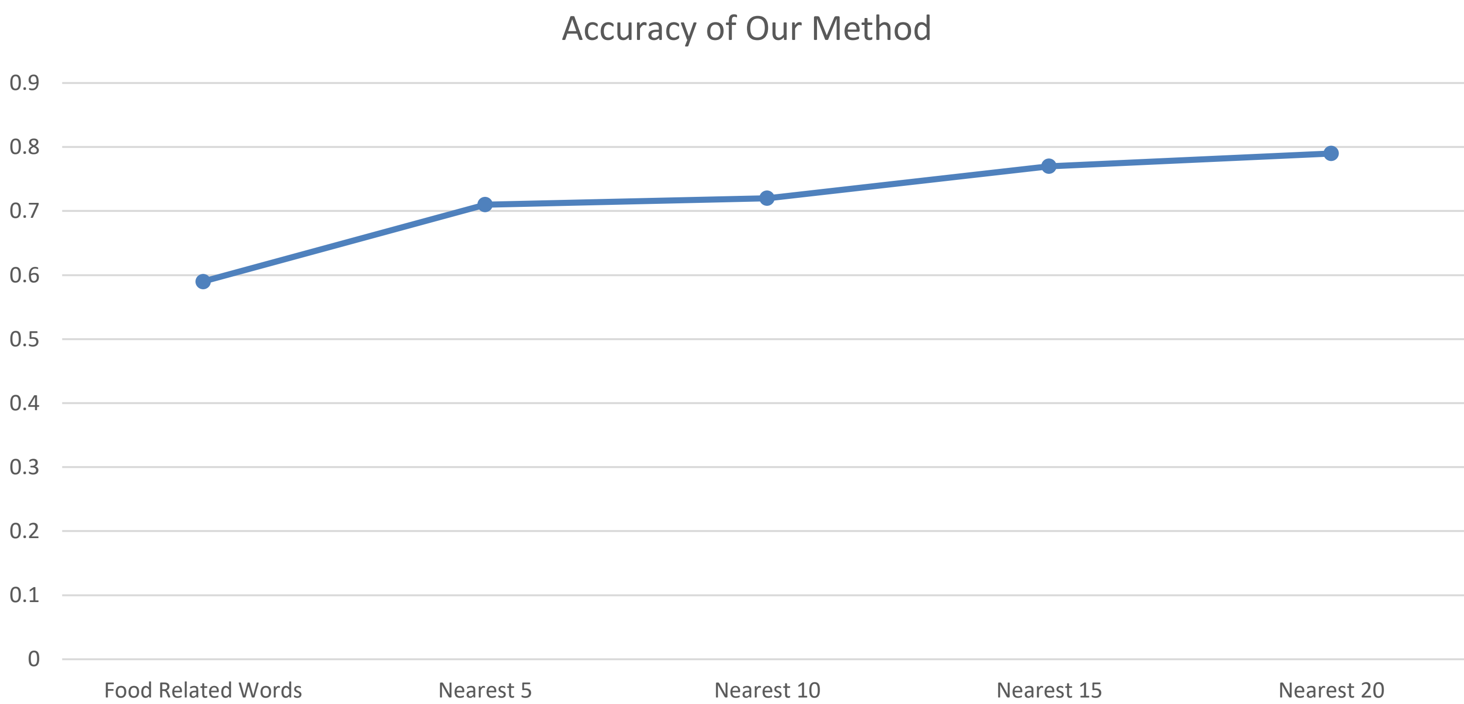
- An example of lexical extension

Classification

- Classification Task
  - 51 states (including Washington D. C.) in the United States
  - Each state is regarded as a single unit
  - Label whether a state is “overweight” or not by comparing with the overweight rate of the national median
- Food related words and extended words by Word2Vec are regarded as features
- TF-IDF normalization for features
- Multiple Layer Perceptron (MLP) Classifier
  - http://deeplearning4j.org
  - “A multilayer perceptron is a logistic regressor where instead of feeding the input to the logistic regression you insert a intermediate layer, called the hidden layer, that has a nonlinear activation function (usually tanh or sigmoid)“
  - Can handle numeric features

Evaluation

- Ten-folder evaluation, get average accuracy



Result

Model	Overweight Accuracy(%)
Majority Baseline	50
Food + LDA + SVM (Fried et al., 2014)	69
Food + Word2Vec + MLP	79

Findings and Discussions

- GIGO (garbage in, garbage out), good features are really significant
- Word2Vec is useful for improving the performance
- Our model can perform better than Food + LDA + SVM

Conclusions and Future Directions

- Our model can well classify overweight on state-level, which provides us with an automatic and scalable methodology to deal with large amount of the social media
- In the future, we will improve the performance of our model by extending our lexical words and at the same time set up thresholds to refine the extension
- We are considering moving from state-level to individual-level

References

Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, “Ranking with recursive neural networks and its application to multi-document summarization,” *Proc. Twenty-Ninth AAAI Conf. Artif. Intell.*, pp. 2153–2159, 2015.

C. Chew and G. Eysenbach, “SARS revisited: Managing ‘outbreaks’ with ‘communications’,” *PLoS One*, vol. 5, no. 11, 2010.

L. Dong, F. Wei, M. Zhou, and K. Xu, “Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis,” *Proc. Twenty-Eighth AAAI Conf. Artif. Intell.*, pp. 1537–1543, 2014.

D. Fried, M. Surdeanu, S. Kobourov, M. Hingle, and D. Bell, “Analyzing the language of food on social media,” *Proc. 2014 IEEE Int. Conf. Big Data (IEEE BigData)*, pp. 778–783, 2014.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data.,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” *Proc. Twenty-Ninth AAAI Conf. Artif. Intell.*, pp. 2267–2273, 2015.

M. J. Paul and M. Dredze, “You are what you tweet: analyzing Twitter for public health,” *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media*, pp. 265–272, 2011.

H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. P. Seligman, and L. Ungar, “Characterizing geographic variation in well-being using tweets,” *Proc. Seventh Int. AAAI Conf. Weblogs Soc. Media*, pp. 583–591, 2013.

E. Yom-Tov, D. Borsa, I. J. Cox, and R. A. McKendry, “Detecting disease outbreaks in mass gatherings using internet data,” *J. Med. Internet Res.*, vol. 16, no. 6, 2014.