A pilot cognitive computing system to understand immunization programs

Marija Stanojevic
Temple University
marija.stanojevic@temple.edu

ABSTRACT

While quantitative data results are used to describe immunization programs, textual data results are much harder to obtain. This may lead to missed opportunities to improve program performance. This paper presents early results from the development of a pilot cognitive computing system designed to systematically analyze a large volume of textual data relevant to immunization programs.

AUDIENCE

Areas: Data Science, Artificial Intelligence, Natural Language Processing, Health Informatics
Intermediate talk

INTRODUCTION

The goal of immunization programs is to maintain and improve vaccination coverage to prevent diseases. For this purpose there are many studies analyzing quantitative data. Qualitative data provide important contextual information necessary to better understand the quantitative data. However, routinely analyzing text data can be labor intensive. This paper describes challenges and solutions in the development of a pilot cognitive computing system to support immunization programs.

In order to obtain accurate lexicons for use by the cognitive computing system, a large quantity of immunization program specific formal and informal data was identified and appropriate cleaning processes were required. Word2vec [1, 2], doc2vec [3], glove [4] and word-topic mixture (WTM) [5] algorithms were adapted to create lexicons. Each lexicon was manually evaluated by domain experts and scored in automatic evaluation.

METHODOLOGY

Data description

Qualitative data were collected from a variety of sources that were categorized as formal and informal. Sources of formal data included vaccine-related websites, scientific documents, journals and books, and current state vaccination-related laws. Sources of informal data included Twitter, online forums and news, and social media feeds. Formal data were characterized by well-structured sentences and paragraphs; however, it could reference external documents and have multiple topics (e.g., a legal statute that covers more than the immunization program). Informal data were characterized by jargon terms, misspelled words, slang language, and metaphors but was generally focused on one topic.

The original objective of the formal data analysis was to show if changes in jurisdictional laws were associated with changes in vaccination coverage and exemptions. However, available jurisdictional laws contained very little data (30 MB). Additionally, this data referenced external documents that were not readily available. The formal data analysis objectives were revised and formal data were collected from additional sources. Experts provided a list of immunization-related websites and journals. General scrapers and parsers were developed to extract useful information from those resources. The data crawling and parsing process was automatized, so that system could work with the most recent data. In total, 1GB of formal data were collected. A limitation of the formal data was that spatio-temporal information for the text or previous versions of text were not always readily available.

Informal data (2.4GB) posted from November 2016 through February 2018 were collected through searching social media data with an online service, Sysomos. The preliminary results were retrieved using multiple queries containing common immunization terms, such as “varicella”, “vaccination”, and “mmr”. However, more than 2/3 of data were not related to vaccination because some words have double meaning. For example, “mmr” is used as both the abbreviation for measles, mumps, and rubella vaccine by immunization programs and the abbreviation for Match Making Rank, a rating system used in online gaming. In addition, the same text was retrieved through multiple queries. The data from the multiple queries were discarded. Instead, one complex logic query was created to retrieve the informal data. In cases where certain words had multiple meanings, results were restricted using logical operators, so that only results that contain two words related to immunization were retained.

Data preprocessing

Only textual information were used for creating lexicons, even though quantitative data such as location, time, and other features were available. Text was cleaned from links, user-names, special characters (except dash and apostrophe symbols), and numbers and then split into words/tokens. Stop-words were removed from tokens. Jargon terms, connected words, and misspellings were kept in the data to allow routinely used language describing immunization to be used in future analysis.

Lexicon creation

Several algorithms and related versions were used to develop the lexicon: word2vec and its variation doc2vec, glove and WTM. Word2vec algorithm is a small neural network with very little parameters that allows fast learning of vector representation of each word. It has four different versions distinguished by sampling of output (hierarchical softmax or negative sampling) and by learning architecture (continuous bag of words which is learning word from context and skip-gram which is learning context from word). Doc2vec is an upgrade of word2vec algorithm that is also able to learn vectors of documents. It has two versions based on architecture that can learn word from paragraph and words (PV-DM) or can learn words from paragraph (PV-DBOW). All of those six versions were evaluated on both formal and informal datasets using different number of vector sizes. Evaluation was done by subject matter experts and using a standard benchmark for evaluation of language representation [1].

Glove algorithm builds matrix of words co-occurrences X_{ij} and then optimizes cost function:

$$\hat{J}=\sum_{i,j} f\left(X_{ij}\right)\left(w_i^T \bar{w}_j - \log X_{ij}\right)^2$$

where $w \in R^d$ are word vectors and $\bar{w} \in R^d$ are context word vectors and

$$f(x)=\begin{cases} \left(x/x_{max}\right)^{3/4}, ifx<x_{max} \\ 1, otherwise \end{cases}.$$

This function prevents influence from large co-occurrence of common words (e.g., “on the”, “at that”, “he is”). The glove algorithm generally shows better results than word2vec [4], but we found that algorithm was not always stable.

WTM model is a purely statistical model which combines ideas from latent Dirichlet allocation (LDA) and word2vec algorithms to learn both representation of words and topics simultaneously. The WTM model generally shows better results than word2vec [5], but it is more complex model and slower at producing output.

Cognitive computing system

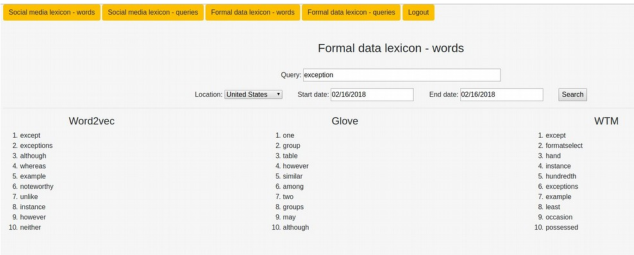
The cognitive computing system uses best scored version of word2vec/doc2vec, glove and WTM algorithm, showing ten most similar word/paragraph results to a given query for each of the algorithms together with their similarity scores. Paragraphs and query vectors were calculated as average of vectors of words contained in each paragraph in formal data or for each document (tweet, comment, news) in informal data. Since many programs will be evaluated by individual jurisdictions, searches based on text data from a single jurisdiction were implemented. Additionally, searches for date range for the informal data were implemented to reflect the changes in language over the period, especially in cases of disease outbreaks. Moreover, system scores jurisdictions for a search query by averaging scores of first ten most similar paragraphs to that query. This distinguishes jurisdictions that are closely related to query term from those that are not, giving the score of similarity. Finally, the system determined if the document (tweet, comment, news) from the informal data was positively or negatively related to vaccination using the trained classification model.

OUTCOMES

Evaluation of word2vec/doc2vec algorithms showed the best outcomes were given by continuous bag of word with negative sampling followed by PV-DM on both datasets. References used different vectors lengths – number of features to describe words. For versions of word2vec algorithm, vectors of length 50 gave best results, even though recommended setting in the references are different. Change in vector size for glove algorithm wasn't that crucial as in word2vec algorithms.

The cognitive computing system displays the results of the best version of word2vec, glove, and WTM algorithms results in parallel. Analysis showed that different aspects of similarity are brought to the top, which is useful in certain applications. The word2vec result was usually better in understanding relationships between query and paragraph, while glove algorithm results were more influenced by frequency of occurrence of the most significant words in query and paragraph.

This system is a web service that communicates with algorithms to get results for queries and to understand if results are positively or negatively related to vaccination. Example of a screen of interface is given in Figure 1 and **functionalities of the cognitive system will be demonstrated in the presentation**.


Social media lexicon - words
Social media lexicon - queries
Formal data lexicon - words
Formal data lexicon - queries
Logout
Search

Formal data lexicon - words

Query: exception

Location: United States Start date: 02/16/2018 End date: 02/16/2018

Word2vec	Glove	WTM
1. except	1. one	1. except
2. exceptions	2. group	2. formatselect
3. although	3. table	3. hand
4. whereas	4. however	4. instance
5. example	5. similar	5. hundredth
6. noteworthy	6. among	6. exceptions
7. unlike	7. two	7. example
8. instance	8. groups	8. least
9. however	9. may	9. occasion
10. neither	10. although	10. possessed

Illustration 1: Figure 1: Example of lexicon search for words, formal dataset

CONCLUSION

Appropriate word and paragraph representation were achieved through a process of text collection and cleaning followed by adapting state-of-the-art algorithms for text representation. The algorithms used brought different aspects of similarity and had different memory and time usage properties. This variety of results is a great starting point for further analysis of textual results, some of which will be addressed in this project, including counting repetition of informal documents and understanding the relation between documents. The cognitive computing system can stay relevant as long as it is able to collect new data and use the data to update models.

The focus of future work is on additional analysis of immunization data to improve development of the cognitive computing system and improve the analytic capacity to support immunization programs.

PARTICIPATION STATEMENT

I will attend the conference if the paper is accepted.

ACKNOWLEDGMENT

This research was supported by the Centers for Disease Control through project Pilot to Assess Cognitive Computing to Analyze Immunization Program Data.

DISCLAIMER

The findings and conclusions in this report are those of the authors and do not necessarily represent the official opinion of the Centers for Disease Control and Prevention.

CONTRIBUTORS

Contributers to this project are:

- Fang Zhou and Zoran Obradovic from Temple University;
- Sarah Ball, William Campbell, Alison Thaung and Sarah Hamad from Abt Associates;
- Stacie Greby, Alexandra Bhatti, Allison Fisher, Yoonjae Kang, Cynthia Knighton and Pamela Srivastava from Centers for Disease Control.

REFERENCES

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[2] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[3] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

[4] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[5] Fu, X., Wang, T., Li, J., Yu, C., & Liu, W. (2016, November). Improving Distributed Word Representation and Topic Model by Word-Topic Mixture Model. In Asian Conference on Machine Learning (pp. 190-205).

BIO

Marija Stanojevic is a PhD student and Research Assistant at Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, US. Her main research focus is language representation and understanding. Marija's interests are natural language processing, machine learning, data mining, statistics, deep learning and education. Before coming to Temple, she worked for year and half as a data software engineer at Arbor Education Partners, UK edTech company. Marija finished master studies in signal processing and bachelor in software engineering at University of Belgrade, Serbia.

She is a member of ACM and ACM-W organisations and she volunteers at TechGirlz delivering computer science workshops to middle school girls to get them interested in different areas of this field. She is a winner of Temple's most prestigious, presidential, fellowship for PhD students. While Marija was an undergraduate student she worked on many projects and in different European teams as part of Board of European Students of Technology. As high school student she founded International Science Festival "Science is not Boogeyman" that is having its 10th edition this year.