Challenges and opportunities for public health made possible by advances in natural language processing

Oliver Baclic1*, Matthew Tunis1, Kelsey Young1, Coraline Doan2, Howard Swerdfeger2, Justin Schonfeld3*

Abstract

Natural language processing (NLP) is a subfield of artificial intelligence devoted to understanding and generation of language. The recent advances in NLP technologies are enabling rapid analysis of vast amounts of text, thereby creating opportunities for health research and evidence-informed decision making. The analysis and data extraction from scientific literature, technical reports, health records, social media, surveys, registries and other documents can support core public health functions including the enhancement of existing surveillance systems (e.g. through faster identification of diseases and risk factors/at-risk populations), disease prevention strategies (e.g. through more efficient evaluation of the safety and effectiveness of interventions) and health promotion efforts (e.g. by providing the ability to obtain expert-level answers to any health related question). NLP is emerging as an important tool that can assist public health authorities in decreasing the burden of health inequality/inequity in the population. The purpose of this paper is to provide some notable examples of both the potential applications and challenges of NLP use in public health.

Introduction

There is a growing interest in deploying artificial intelligence (AI) strategies to achieve public health outcomes, particularly in response to the global coronavirus disease 2019 (COVID-19) pandemic where novel datasets, surveillance tools and models are emerging very quickly.

The objective of this manuscript is to provide a framework for considering natural language processing (NLP) approaches to public health based on historical applications. This overview includes a brief introduction to AI and NLP, suggests opportunities where NLP can be applied to public health problems and describes the challenges of applying NLP in a public health context. Particular articles were chosen to emphasize the breadth of potential applications for NLP in public health as well as the not inconsiderable challenges and risks inherent in incorporating AI/NLP in public health analysis and decision support.

Artificial intelligence and natural language processing

AI research has produced models that can interpret a radiograph (1,2), detect irregular heartbeats using a smartwatch (3), automatically identify reports of infectious disease in the media (4), ascertain cardiovascular risk factors from retinal images (5) and find new targets for existing medications (6,7). The success of these models is built from training on hundreds, thousands and sometimes millions of controlled, labelled and structured data points (8). The capacity of AI to provide constant, tireless and rapid analyses of data offers the potential to transform society’s approach to promoting health and preventing and managing diseases. AI systems have the potential to “read” and triage all of the approximately 1.3 million research articles indexed by PubMed each year (9); “examine” comments from 1.5 billion Facebook users or “monitor” 500 million tweets of people struggling with mental illness on a daily basis, foodborne illness or the flu (10,11); and simultaneously interact with each and every person seeking answers to their health questions, concerns, problems and challenges (12).
NLP is a subfield of AI that is devoted to developing algorithms and building models capable of using language in the same way humans do (13). It is routinely used in virtual assistants like “Siri” and “Alexa” or in Google searches and translations. NLP provides the ability to analyze and extract information from unstructured sources, automate question answering and conduct sentiment analysis and text summarization (8). With natural language (communication) being the primary means of knowledge collection and exchange in public health and medicine, NLP is the key to unlocking the potential of AI in biomedical sciences.

Most modern NLP platforms are built on models refined through machine learning techniques (14,15). Machine learning techniques are based on four components: a model; data; a loss function, which is a measure of how well the model fits the data; and an algorithm for training (improving) the model (16). Recent breakthroughs in these areas have led to vastly improved NLP models that are powered by deep learning, a subfield of machine learning (17).

Innovation in the different types of models, such as recurrent neural network-based models (RNN), convolutional neural network-based models (CNN) and attention-based models, has allowed modern NLP systems to capture and model more complex linguistic relationships and concepts than simple word presence (i.e. keyword search) (18). This effort has been aided by vector-embedding approaches to preprocess the data that encode words before feeding them into a model. These approaches recognize that words exist in context (e.g. the meanings of “patient,” “shot” and “virus” vary depending on context) and treat them as points in a conceptual space rather than isolated entities. The performance of the models has also been improved by the advent of transfer learning, that is, taking a model trained to perform one task and using it as the starting model for training on a related task. Hardware advancements and increases in freely available annotated datasets have also boosted the performance of NLP models. New evaluation tools and benchmarks, such as GLUE, superglue and BioASQ, are helping to broaden our understanding of the type and scope of information these new models can capture (19–21).

**Opportunities**

Public health aims to achieve optimal health outcomes within and across different populations, primarily by developing and implementing interventions that target modifiable causes of poor health (22–26). Success depends on the ability to effectively quantify the burden of disease or disease risk factors in the population and subsequently identify groups that are disproportionately affected or at-risk; identify best practices (i.e. optimal prevention or therapeutic strategies); and measure outcomes (27). This evidence-informed model of decision making is best represented by the PICO concept (patient/problem, intervention/exposure, comparison, outcome). PICO provides an optimal knowledge identification strategy to frame and answer specific clinical or public health questions (28). Evidence-informed decision making is typically founded on the comprehensive and systematic review and synthesis of data in accordance with the PICO framework elements.

Today, information is being produced and published (e.g. scientific literature, technical reports, health records, social media, surveys, registries and other documents) at unprecedented rates. By providing the ability to rapidly analyze large amounts of unstructured or semistructured text, NLP has opened up immense opportunities for text-based research and evidence-informed decision making (29–34). NLP is emerging as a potentially powerful tool for supporting the rapid identification of populations, interventions and outcomes of interest that are required for disease surveillance, disease prevention and health promotion. For example, the use of NLP platforms that are able to detect particular features of individuals (population/problem, e.g. a medical condition or a predisposing biological, behavioural, environmental or socioeconomic risk factor) in unstructured medical records or social media text can be used to enhance existing surveillance systems with real-world evidence. One recent study demonstrated the ability of NLP methods to predict the presence of depression prior to its appearance in the medical record (35). The ability to conduct real-time text mining of scientific publications for a particular PICO concept provides opportunities for decision makers to rapidly provide recommendations on disease prevention or management that are informed by the most current body of evidence when timely guidance is essential, such as during an outbreak. NLP-powered question-answering platforms and chatbots also carry the potential to improve health promotion activities by engaging individuals and providing personalized support or advice. Table 1 provides examples of potential applications of NLP in public health that have demonstrated at least some success.

**Challenges**

Despite the recent advances, barriers to widespread use of NLP technologies remain.

Similar to other AI techniques, NLP is highly dependent on the availability, quality and nature of the training data (72). Access and availability of appropriately annotated datasets (to make effective use of supervised or semi-supervised learning) are fundamental for training and implementing robust NLP models. For example, the development and use of algorithms that are able to conduct a systematic synthesis of published research on a particular topic or an analysis and data extraction from electronic health records requires unrestricted access to publisher or primary care/hospital databases. While the number of freely accessible biomedical datasets and pre-trained models has been increasing in recent years, the availability of those dealing with public health concepts remains limited (73).
The ability to de-bias data (i.e. by providing the ability to inspect, explain and ethically adjust data) represents another major consideration for the training and use of NLP models in public health settings. Failing to account for biases in the development (e.g. data annotation), deployment (e.g. use of pre-trained platforms) and evaluation of NLP models could compromise the model outputs and reinforce existing health inequity (74). However, it is important to note that even when datasets and evaluations are adjusted for biases, this does not guarantee an equal impact across morally relevant strata. For example, use of health data available through social media platforms must take into account the specific age and socioeconomic groups that use them. A monitoring system trained on data from Facebook is likely to be biased towards health data and linguistic quirks specific to a population older than one trained on data from Snapchat (75). Recently many model agnostic tools have been developed to assess and correct unfairness in machine learning models in accordance with the efforts by the government and academic communities to define unacceptable AI development (76–81).

Currently, one of the biggest hurdles for further development of NLP systems in public health is limited data access (82,83). Within Canada, health data are generally controlled regionally and, due to security and confidentiality concerns, there is reluctance to provide unhindered access to these systems and their integration with other datasets (e.g. data linkage). There have also been challenges with public perception of privacy and data access. A recent survey of social media users found that the majority considered analysis of their social media data to identify mental health issues “intrusive and exposing” and they would not consent to this (84).

Before key NLP public health activities can be realized at scale, such as the real-time analysis of national disease trends, jurisdictions will need to jointly determine a reasonable scope and access to public health–relevant data sources (e.g. health record and administrative data). In order to prevent privacy violations and data misuse, future applications of NLP in the analysis of personal health data are contingent on the ability to embed differential privacy into models (85), both during training and postdeployment. Access to important data is also limited through the current methods for accessing full text publications. Realization of fully automated PICO-specific knowledge extraction and synthesis will require unrestricted access to journal databases or new models of data storage (86).

Finally, as with any new technology, consideration must be given to assessment and evaluation of NLP models to ensure that they are working as intended and keeping in pace with society’s changing ethical views. These NLP technologies need to be assessed to ensure they are functioning as expected and account for bias (87). Although today many approaches are posting equivalent or better-than-human scores on textual analysis tasks, it is important not to equate high scores with true language understanding. It is, however, equally important not to view

Table 1: Examples of existing and potential applications of natural language processing in public health

<table>
<thead>
<tr>
<th>Type of activity</th>
<th>Public health objective</th>
<th>Example of NLP use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of at-risk populations or conditions</td>
<td>To continuously measure the incidence and prevalence of diseases and disease risk factors (i.e. surveillance)</td>
<td>Analysis of unstructured or semistructured text from electronic health records or social media (36–42)</td>
</tr>
<tr>
<td>of interest</td>
<td>To identify vulnerable and at-risk populations</td>
<td>Analysis of risk behaviours using social media (43–45)</td>
</tr>
<tr>
<td>Identification of health interventions</td>
<td>To develop optimal recommendations/interventions</td>
<td>Automated systematic review and analysis of the information contained in scientific publications and unpublished data (46–50)</td>
</tr>
<tr>
<td></td>
<td>To identify best practices</td>
<td>Identification of promising public health interventions through analysis of online grey and peer reviewed literature (51)</td>
</tr>
<tr>
<td>Identification of health outcomes using real-world</td>
<td>To evaluate the benefits of health interventions</td>
<td>Analysis of unstructured or semistructured text from electronic health records, online media and publications to determine the impact of public health recommendations and interventions (52,53)</td>
</tr>
<tr>
<td>evidence</td>
<td>To identify unintended adverse outcomes related to interventions</td>
<td>Analysis of unstructured or semistructured text from electronic health records, social media and publications to identify potential adverse events of interventions (54–58)</td>
</tr>
<tr>
<td>Knowledge generation and translation</td>
<td>To support public health research</td>
<td>Analysis and extraction of information from electronic health records and scientific publications for knowledge generation (59–62)</td>
</tr>
<tr>
<td></td>
<td>To support evidence-informed decision making</td>
<td>Use of chatbots, question/answer systems and text summarizers to provide personalized information to individuals seeking advice to improve their health and prevent disease (63–65)</td>
</tr>
<tr>
<td>Environmental scanning and situational awareness</td>
<td>To conduct public health risk assessments and provide situational awareness</td>
<td>Analysis of online content for real-time critical event detection and mitigation (66–70)</td>
</tr>
<tr>
<td></td>
<td>To monitor activities that may have an impact on public health decision making</td>
<td>Analysis of decisions of international and national stakeholders (71)</td>
</tr>
</tbody>
</table>

Abbreviation: NLP, natural language processing
a lack of true language understanding as a lack of usefulness. Models with a “relatively poor” depth of understanding can still be highly effective at information extraction, classification and prediction tasks, particularly with the increasing availability of labelled data.

**Natural language processing and the coronavirus disease 2019 (COVID-19)**

With the emergence of the COVID-19, NLP has taken a prominent role in the outbreak response efforts (88,89). NLP has been rapidly employed to analyze the vast quantity of textual information that has been made available through unrestricted access to peer-review journals, preprints and digital media (90). NLP has been widely used to support the medical and scientific communities in finding answers to key research questions, summarization of evidence, question answering, tracking misinformation and monitoring of population sentiment (91–97).

**Conclusion**

NLP is creating extraordinary opportunities to improve evidence-informed decision making in public health. We anticipate that broader applications of NLP will lead to the creation of more efficient surveillance systems that are able to identify diseases and at-risk conditions in real time. Similarly, with an ability to analyze and synthesize large volumes of information almost instantaneously, NLP is expected to facilitate targeted health promotion and disease prevention activities, potentially leading to population-wide disease reduction and greater health equity. However, these opportunities are not without risks: biased models, biased data, loss of data privacy and the need to maintain and update models to reflect the evolving language and context of public communication are all existing challenges that will need to be addressed. We encourage the public health and computer science communities to collaborate in order to mitigate these risks, ensure that public health practice does not fall behind in these technologies or miss opportunities for health promotion and disease surveillance and prevention in this rapidly evolving landscape.

**Authors’ statement**

OB — Writing – original draft, review & editing and conceptualization
MT — Writing – original draft, review & editing and conceptualization
KY — Writing – review & editing, and conceptualization
CD — Writing – review & editing
HS — Writing – review & editing
JS — Writing – original draft, review & editing and conceptualization

**Conflict of interest**
None.

**Acknowledgements**
We thank J Nash and J Robertson who were kind enough to offer feedback and suggestions.

**Funding**
This work is supported by the Public Health Agency of Canada. The research undertaken by JS was funded by the Canadian federal government’s Genomic Research and Development Initiative.

**References**


49. Norman C, Leeflang M, Spijker R, Kanoulas E, Névéol A. A distantly supervised dataset for automated data extraction from diagnostic studies. ACL Workshop on Biomedical Natural Language Processing, Florence (IT): 2019 Aug. DOI


