Cross Disciplinary Consultancy to Bridge Public Health Technical Needs and Analytic Developers: Negation Detection Use Case

Mike Conway[1], Danielle Mowery[1,2], Amy Ising[3], Sumithra Velupillai[4,5], Son Doan[6], Julia Gunn[7], Michael Donovan[7], Caleb Wiedeman[8], Lance Ballester[9], Karl Soetebier[9], Catherine Tong[10], Howard Burkom[11]

1. Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, United States

2. Informatics, Decision-Enhancement, and Analytical Sciences Center (IDEAS 2.0), Veterans Affairs, Salt Lake City Health Care System, Salt Lake City, Utah, United States

3. Department of Emergency Medicine, University of North Carolina Chapel Hill, North Carolina, United States

4. Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom

5. School of Electrical Engineering & Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

6. Medical Informatics, Kaiser Permanente Southern California, San Diego, California, United States

7. Boston Public Health Commission, Boston, Massachusetts, United States

8. Tennessee Department of Health, Nashville, Tennessee, United States

9. Georgia Department of Public Health, Atlanta, Georgia, United States

10. International Society for Disease Surveillance, Brighton, Massachusetts, United States

11. Applied Physics Laboratory, Johns Hopkins University, Baltimore, Maryland, United States

Abstract

This paper describes a continuing initiative of the International Society for Disease Surveillance designed to bring together public health practitioners and analytics solution developers from both academia and industry. Funded by the Defense Threat Reduction Agency, a series of consultancies have been conducted on a range of topics of pressing concern to public health (e.g. developing methods to enhance prediction of asthma exacerbation, developing tools for asyndromic surveillance from chief complaints). The topic of this final consultancy, conducted at the University of Utah in January 2017, is focused on defining a

roadmap for the development of algorithms, tools, and datasets for improving the capabilities of text processing algorithms to identify negated terms (i.e. *negation detection*) in free-text chief complaints and triage reports.

Keywords: negation detection, natural language processing, syndromic surveillance, chief complaints

Correspondence: mike.conway@utah.edu

DOI: 10.5210/ojphi.v10i2.8944

1. Introduction

Despite considerable effort since the turn of the century to develop Natural Language Processing (NLP) methods and tools for public health surveillance, few standardised methods have emerged. Those methods that have emerged (e.g. the NegEx algorithm [1]) are confined to local implementations with customised solutions. Furthermore, agreement on international standards — even between countries as geographically proximal and culturally similar as the United States and Canada — has proved elusive. Important reasons for this lack of progress include (a) limited shareable datasets for developing and testing methods (b) jurisdictional data silos, and (c) the gap between resource-constrained public health practitioners and technical solution developers, typically university researchers and industry developers.

To address these three problems, the *International Society for Disease Surveillance* (ISDS) launched a Technical Conventions Committee in 2013, tasked with collecting and curating surveillance-related use cases from public health stakeholders at the local, state, and federal levels in the United States, and generating detailed requirement templates and datasets to support these use cases, with the goal of developing and disseminating best practice to the public health community. In 2014, ISDS was awarded a three-year contract by the Defense Threat Reduction Agency (DTRA) (Analytic Solutions for Real-Time Biosurveillance Project) with the aim of bringing together public health practitioners and solution developers with the overarching goal of supporting DTRA's Biosurveillance Ecosystem (BSVE) [2,3]. BSVE is a web-based, open-source, cloud-hosted dashboard for managing, visualising, and analysing disparate health and non-health related data to support public health surveillance and situational awareness.

In this paper, we describe the process and results of a consultancy involving four types of public health stakeholders, including (1) representatives from public health departments (local, state, federal), (2) university researchers focused on computational methods for public health surveillance, (3) members of public health oriented non-governmental organisations, and (4) industry representatives, all interested in developing validated, standardised and portable resources (meth-

ods and data sets) for negation detection in clinical text used for public health surveillance. As will be described more fully below, accurate negation detection is a vital step required for increasing the reliability of text-based syndromic surveillance methods.

2. Materials and Methods: Consultancy

Under the ISDS DTRA grant, five use case consultancies were conducted, as summarised in Table 1. Hosts of the first four consultancies found these events highly successful for communicating cross-disciplinary needs and analysing technological requirements [2,3]. The main objective of this grant was to develop an understanding of needs and research capabilities to support accurate negation detection for public health. The adaptation and implementation of existing methods in the host surveillance systems were beyond the scope of the grant and remain a challenge.

Interpreting "methods" as the elicitation and dialogic processes of the consultancy and "materials" as the composition of participants, this negation consultancy differed qualitatively from the prior consultancies under the ISDS DTRA grant. For example, prior events were driven by a single health department (except for the second consultancy hosted by the United States Department of Defense) seeking practical tools to meet specific, routine surveillance needs. In contrast, this final use case-based consultancy was hosted, not by a health-monitoring organisation, but by a large, prolific, and well-established informatics department — the Department of Biomedical Informatics at the University of Utah – with connections to similar academic, governmental, and industrial research groups. Knowing the interest of many health departments in improving and broadening their use of NLP for medical data, the grant's Advisory Group chose an academic hosting approach to attract a variety of related expertise. Moreover, instead of motivation from a single health department, needs were presented by three state (Tennessee, Georgia, North Carolina) and two local (King County WA, Boston MA) health departments with a range of surveillance capabilities and experience in usage of free-text data.

Table 1: Dates, hosts, and subject matter of the five DTRA-funded consultancies conducted by ISDS

Date	Host	Subject
Jun 9-10, 2015	North Carolina Division of Public Health	Asyndromic Cluster Detection in ED CC Text
Oct 29-30, 2015	U.S. Department of Defense	Predictive Models for Infectious Disease
Mar 30-31, 2016	Boston Public Health Commission	Estimating Risk of Asthma Exacerbations
Jun 14-15, 2016	Arizona Department Health	Risk Mapping for Incidence of Arboviral Disease
Jan 19-20, 2017	University of Utah	Negation Processing in Free-Text Analysis

The primary need for text processing is classification of records — for most users so far, Emergency Department chief complaint text — into indicator bins in order that the size of each bin (i.e. the volume of classified chief complaints) may be monitored on a weekly, daily, or more frequent basis. Most of the free-text data already used by public health departments are chief complaint or reason-for-visit fields. Entries in these fields are often fewer than six words and may lack verbs or punctuation (see Table 2 for example chief complaints). Health department users employ a limited range of standardised tools for classification and these tools are often neither well understood by in-house staff (i.e. they are "black box" algorithms) nor optimised for local application. Key application questions underlying the consultancy were:

Table 2: Example Emergency Department chief complaints

cva	DIFGF BREATHING
Chest pain since last night, weakness, abdominal pain	Emesis
Infection on the power port	crisis evaluation
Pt came for Wt and BP ck to start phentermine	follow up
Patient is here for cold symptoms, hx of CHF	Fever
fever,abd pain/from uc	facial pain
annual exam; right leg concern- swelling and pain	headache
alcohol intoxication	cough
spasms	psychosis
neck pain	fall

- Can/should classification tools already available be used more effectively, updated, or replaced to improve sensitivity?
- How effective are more advanced tools for classifying various text-based data sources? (e.g. triage notes, chief complaints, and/or text fields from Emergency Department patient records)
- What additional tools can be implemented and maintained given the scarce resources and limited in-house expertise of health departments?
- While none of the health departments had the goal of eliminating human-in-the-loop routine surveillance activities, trained human expertise is both expensive and limited in supply. Therefore, understanding the costs and benefits involved in interpreting output from the desired tool is essential for effective surveillance system planning.

In a recent ISDS survey and subsequent publication on surveillance research priorities of public health practitioners, one of the top recommendations was the need to develop new "methods to process, categorize, and code unstructured data in electronic health records" [4]. Preparations for the consultancy included an email thread and a shared website for published articles and data samples leading to a structured pre-consultancy call designed to inform participants regarding the purpose of

the consultancy and to align expectations. Then, health department users were requested to provide data samples exemplifying negation issues in the classification process. Presenting developers were asked to explain their underlying ideas, details of method implementation, size and composition of corpora used for evaluation, and classification performance results. As in the prior consultancies, data architecture, networking processes, and multipurpose automated systems were not the focus of the event, but they provided context and constraints for the discussion. Finally, each attending health department was asked for an overview of the data acquisition, preprocessing, analysis, and interpretation steps used in its automated surveillance.

3. Results: Consultancy

The consultancy was held on January 19th & 20th 2017 at the University of Utah's Department of Biomedical Informatics, and consisted of 25 participants. Participants were drawn from various different sectors, with representation from ISDS [2], the Defense Threat Reduction Agency [1], universities and research institutes [5], public health departments [6], the Department of Veterans Affairs [4], non-profit organisations [2], and technology firms [1]. Participants were drawn from a variety of different professional backgrounds, including research scientists, software developers, public health officials, epidemiologists, and analysts. Further, participants were drawn from all major regions of the United States (e.g. South, East Coast, West Coast, Intermountain West) and internationally (United Kingdom & Sweden).

Day 1 of the consultancy was devoted to providing an overview of NLP and current trends in negation detection, including a detailed description of widely used algorithms and tools for the negation detection task. Key questions included: *Should our focus be chief complaints only, or should we widen our scope to Emergency Room triage notes?, How many other NLP tasks (e.g. reliable concept recognition) are necessary to address on the road to improved negation detection?*

With this background established, Day 2 centered on presentations from five different United States local and state health departments (viz: King County WA, Boston MA, North Carolina, Georgia, and Tennessee) on the various approaches to text processing and negation detection across several jurisdictions. We then pivoted and discussed NegEx implementations, discussed the potential utility of social media — particularly Twitter — for syndromic surveillance applications, and watched several demonstrations of syndromic surveillance systems that utilise text processing, including a demonstration of EcoHealth Alliance's surveillance applications, a demonstration of GEOVIZ and ARGUS [6], and a demonstration of topic modelling approaches in the context of the BSVE geospatial syndromic surveillance system.

A post-consultancy survey of participants yielded the following reactions and suggestions:

- Respondents indicated that a greater focus on machine learning approaches (including inviting more machine learning researchers) would have been beneficial.
- Future consultancies could focus explicitly on connecting the public health and computer science communities.

- All participants judged the consultancy to be effective.

Suggested topics for future follow-up meetings included:

- A session on big data text processing for syndromic surveillance in general (i.e. not restricted to the topic of negation detection).
- A session on developing NLP resources for specific problems in syndromic surveillance (e.g. opioid abuse).

4. Materials and Methods: Use Case

4.1. Why free text processing is important for public health

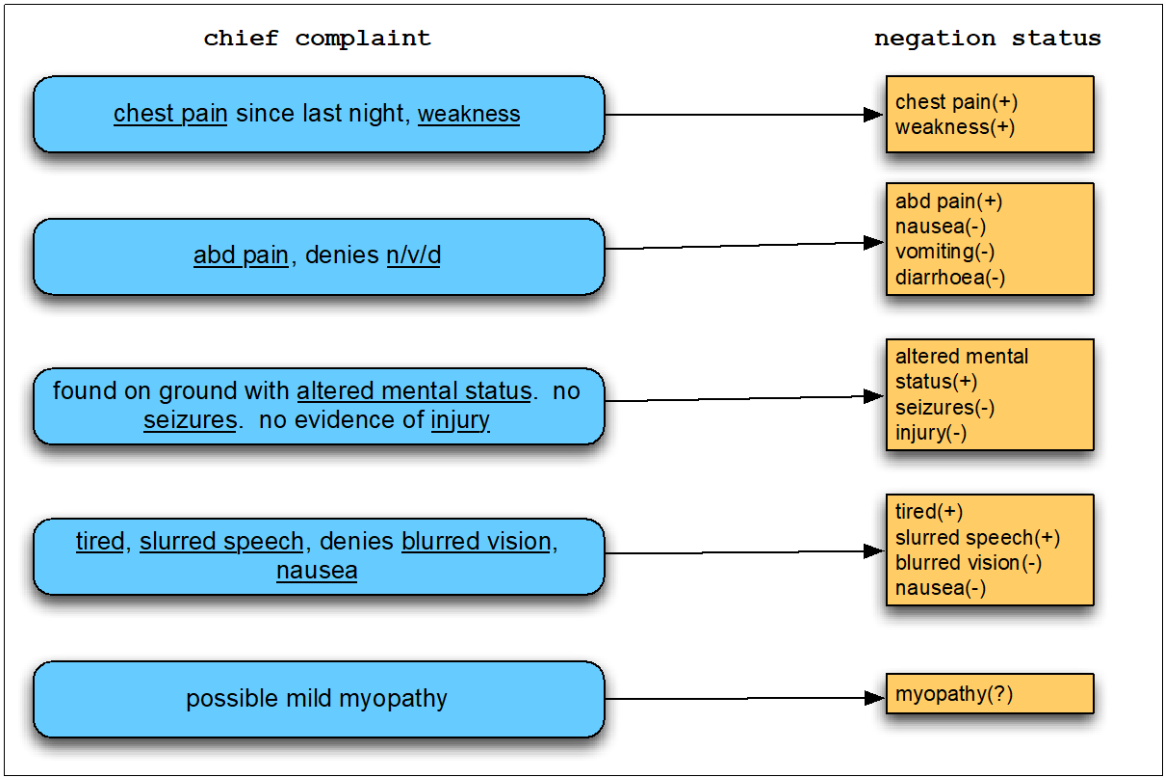
Since the early years of the 21st Century and the widespread adoption of information technology systems in public health departments, free-text chief complaints — short texts that describe a patient's symptoms before a preliminary diagnosis can be made — have come to be a vital resource for syndromic surveillance in the United States context [7]. Traditionally, public health surveillance has relied on the routine filing of reportable disease diagnoses by laboratories or clinicians, sometimes involving considerable delays. In contrast, syndromic surveillance uses data sources that already exist but have not been designed with public health goals in mind. Many systems use Emergency Room-generated triage chief complaints, nearly ubiquitous in the US context and available electronically during, or shortly after a patient encounter. Chief complaint based systems (e.g. NC DETECT [8], EARS [9]) are widely used for customary syndromic surveillance (e.g. annual flu monitoring [10]).

Free-text chief complaints remain a vital resource for syndromic surveillance. However, the widespread adoption of Electronic Health Records (and federal Meaningful Use requirements [5]) has brought changes to the syndromic surveillance practice ecosystem. These changes have included the widespread use of EHR-generated chief complaint "pick lists" (i.e. pre-defined chief complaints that are selected by the user, rather than text strings input by the user at a keyboard), triage note templated text, and triage note free-text (typically much more comprehensive than traditional chief complaints). As will be explained below, a key requirement for a negation detection algorithm is the ability to successfully and accurately process these new and challenging data streams. See Table 2 for examples of the heterogeneity of chief complaint data. Figure 1 provides practical examples of negation detection and concept recognition.

4.2. Impact of Negation

In this section, we first describe current negation detection practices at two public health departments (State of Georgia and City of Boston, MA), before going on to describe important technical and human resource constraints frequently experienced by health departments. Finally, we describe some current approaches to negation detection in the clinical NLP literature.

Figure 1: Chief complaint annotation



4.2.1. Georgia Department of Health

Based on estimates from Georgia data, negation occurs in about 3.5% of incoming non-missing chief complaints, and occurs disproportionally in certain types of syndrome classifications, such as influenza-like illness (ILI), for which negation occurs in up to 6.8% of total visit records. If negation is not handled correctly then syndrome counts can be incorrectly tallied, and because negation occurs disproportionally among syndrome groups, select syndromes will be unequally affected. Not handling negation can then result in incorrect signaling and/or lead to misidentification or misinterpretation of the significance of a syndrome analysis. There are a variety of reasons that off-the-shelf NLP techniques are likely to experience difficulties successfully processing negation in chief complaints. First, the text fields are often short, with a median of 3 words (19 characters), rarely consist of complete sentences, and do not abide by grammatical convention (i.e. they are typically "telegraphic" in character). Second, there are frequent misspellings and many, sometimes seemingly arbitrary, abbreviations. The combination of the two — i.e. misspellings and idiosyncratic abbreviations — in conjunction with little context, makes it difficult to discern the explicit intent of a given chief complaint. Third, there is variability in how the messages are written. One message may be written in well-formed, grammatically correct sentences that directly quotes from patients (example 1 below), while another may be written as a list of symptoms and non-symptoms (example 2 below).

Example 1:

> *"vomitingcough,cold,congestioncough & congestion, vomiting post tussis x 1 week. hed a fever 2 days ago of 102.3. denies fever since. not sleeping well"*

This message has concatenation, poor sentence structure, informative negation on sleeping, and context around the word fever.

Example 2:

> *"llq abd pain, cough - denies n/v/d, fever"*

There are a number of different abbreviations in this message, and little context affecting the words. In both examples there is a question as to whether fever should actually be counted as a symptom or not.

4.2.2. Boston Public Health Commission

In Boston's experience the chief complaint is the most problematic of all the data elements sent though the syndromic surveillance system. The challenges include the aforementioned spelling errors, the use of jargon and abbreviations, and inconsistent documentation by health care workers with various degrees of clinical experience. Boston has developed numerous data management processes to improve the utility and validity of chief complaint processing, including fixing simple spelling errors, removing punctuation and extra spaces, and the use of EMT-P (i.e. Emergency Medical Text Processor) [11] and its attendant software components, the National Library of Medicine's Universal Medical Language System and its Lexical Variant Generator [11,12]. However, one of the main issues with classifying the free-text chief complaint has consistently been negation. Although the size of the chief complaint fields limits the number of negation terms in comparison to those found in longer text fields, ignored negation can result in frequent chief complaint misclassification. The problem very likely worsens for triage notes and for recently adopted "expect" notes, used by Boston hospitals to document patients' clinical issues prior to arrival. Another free-text information source are nursing note fields from electronic disease surveillance systems where negation terms are more frequently found than in chief complaint data. Correctly categorising nursing notes is important for the development of metrics for situational awareness, response management and quality management. A further issue associated with free-text data is the accurate identification of terms expressing uncertainty. This issue is semantically related to negation and involves uncertainty regarding the truth of a particular assertion. For example, the chief complaint "patient thinks he is having a heart attack" is — in the absence of negation and uncertainty detection — interpreted without negation or context analysis as a myocardial infarc- tion. Although not as common as negated terms, uncertainty terms like these pose classification issues. Boston analysed 3.1 million chief complaints for negation terms ("no", "not", or "denies") and uncertainty terms ("thinks", "seems", or "might") from 2013-2016. Negation terms were identified in 8,348

(0.2%) chief complaints and uncertainty terms in 140 chief complaints. In 2016, 1,782 expect notes were received with 542 (30%) containing at least one negation term and 22 (1%) contained at least one uncertainty terms. The changing structure and documentation processes in EHRs are likely to increase the need for data management processes to address linguistic issues. The need for accurate and efficient algorithms for negation detection is growing increasingly urgent.

4.3. Resource Constraints in Public Health Departments

Of the health departments participating in the consultancy, only one was in the process of actively implementing an internal negation solution. Barriers to implementation included limited and siloed information technology resources, differing organisational structures, and inconsistencies in the accepted software standards across health departments. For instance, some health departments have the ability and infrastructure to locally store and analyse incoming information, whereas others pass their data onto outside organisations with little or no manipulation, checking, or tracking done internally. Additionally, there is hesitancy in public health information technology departments to adopt open source solutions without enterprise support. The general consensus that emerged from consultancy participants was that, for sustainable impact, any analytic solution from the group should be able to run successfully in a corporate Windows environment using commodity hardware.

4.4. Current Approaches to Negation Detection

The NegEx algorithm — originally developed in 2001 and the most widely used negation algorithm for clinical text processing [1] — relies on predefined *target concepts* (e.g. “headache”) and a list of *negation terms* that either (a) truly negates the target term (e.g. “**no** headache”) or (b) appear to indicate negation but do not (called *pseudo-negations*), as in “**not** ruled out”. Furthermore, patterns are included to determine whether the negation term *precedes* (e.g. “**no** headache”) or *follows* (“headache **never** developed”) the target concept. If the algorithm finds a target concept, the algorithm then goes on to search for negation terms within a predefined number of surrounding words (typically 5-6). When evaluated on 1,235 target concepts (from 1000 discharge summary sentences), this approach resulted in 84.5% precision/positive predictive value and 77.8% recall/sensitivity [1]. See Figure 2 for an explanation of these metrics.

NegEx has been extended to handle negation scope by defining *conjunction terms* (e.g. “but”, “except”) and other types of modifiers (e.g. experiencer and historicity) in ConText [13], and to add further flexibility in user-defined modifiers, target concepts, and support for document-level assertions in pyConText [14]. NegEx and its extensions have also been used and developed for other use-cases and languages, e.g. Swedish [15], Dutch [16], French [17] and Spanish [18]. A similar, but not identical, lexical approach is presented in Garcelon et al. [19], where regular expressions are defined for French negation expressions along with exclusion rules to address double negation.

Approaches that rely on lexical surface features do not explicitly take *linguistic* relations between words into account, such as the relation between a negation term and the term(s) it syntactically governs (scope). To capture such relations, syntactic parse information can be useful. Sohn et al.

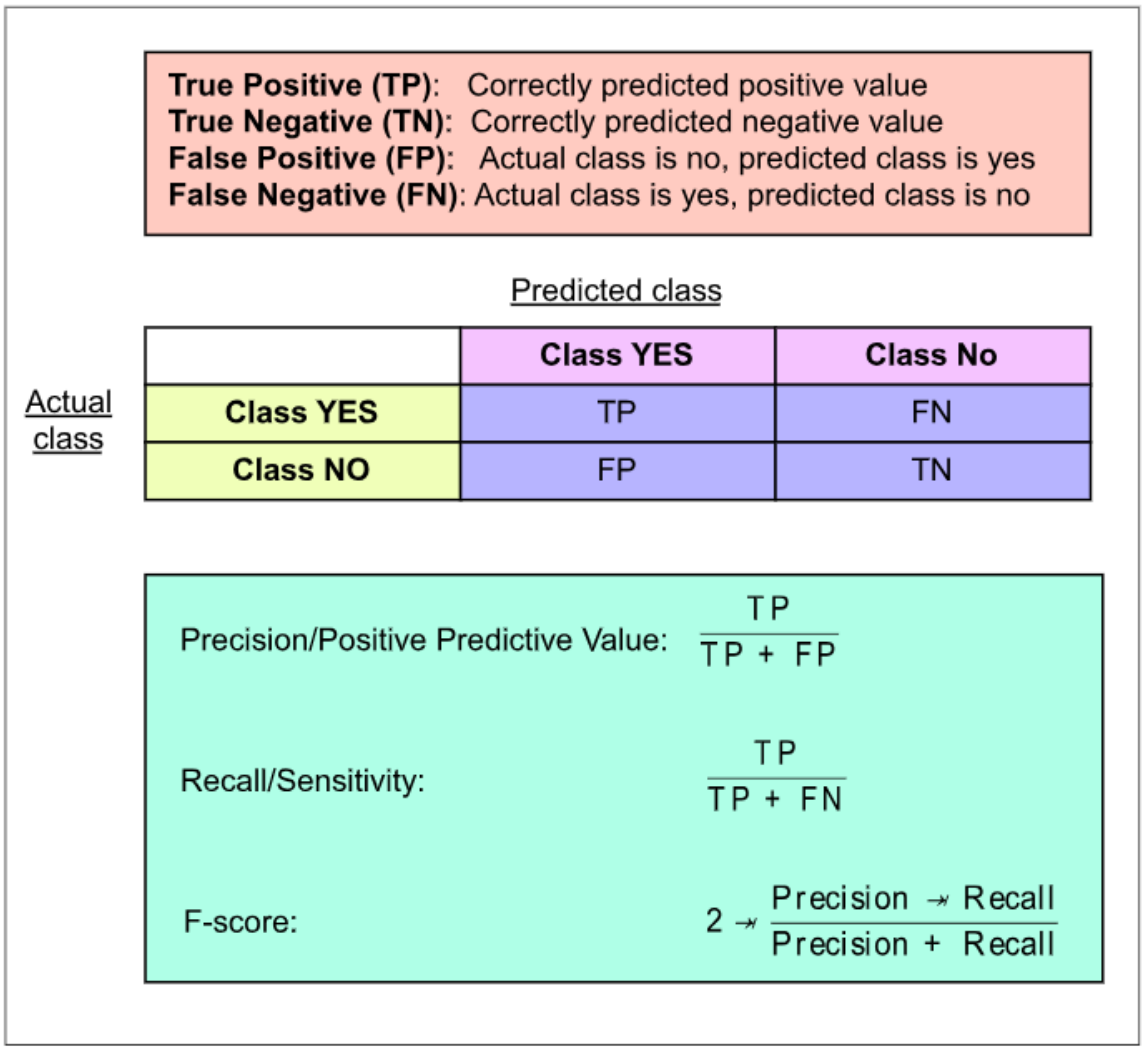
developed an approach which relies on dependency paths between target concepts and negation terms, which reduced Type I errors (false positives) in their evaluation data, reporting very high precision (96.6%) and moderately high recall (73.9%) [20]. Gkotsis et al. used syntactic information from constituency trees (a formalism for representing the grammatical structure of a text) and defines a number of tree traversal operations to determine whether or not a target concept is negated by a negation term (89.4% precision, 94.6% recall) [21].

Lexical approaches have also been combined with syntactic approaches. DEEPEN [22] uses dependency parse tree information to define post-processing rules after first running the NegEx algorithm, an enhancement which results in fewer false positives (89.2-96.6% precision and 73.8-96.3% recall). NegFinder [23] also adds syntactic analysis (context-free grammar) for cases where lexical approaches might fail, such as when the target term and negation term are far from each other (91.8% precision, 95.7% recall). Similarly, Huang et al. combine regular expression matching with grammatical parsing to improve negation scope detection [24]. Tanushi et al. also employed an approach relying on dependency tree information and compared results with NegEx and pyConText on Swedish clinical notes, concluding that results were similar but that the advantage of using syntactic information could be more generalisable [25].

In general, the main difference between lexical, “surface”-based approaches and syntactic approaches is how the scope of negation is detected. Both these approaches rely, however, on rules that are developed manually. With the increased availability of labeled data, such as provided by the 2010 i2b2 challenge [26] for clinical text, approaches that automatically learn patterns (machine learning) have also been applied to this problem with promising results using Support Vector Machines and a variety of features including surrounding contextual, grammatical and syntactic information [27-29].

Common to all these approaches is that the target concept (i.e. the concept that is negated) is defined (more or less explicitly) in advance, and that the negation terms are separate terms or phrases. Recent work includes analysis also of morphological negation (e.g. *im* in impossible) and double negations [30]. However, as is observed by Wu et al. [27], although it is fairly straight- forward to optimise a negation detection approach for a new use case and corpus, developing a generalisable approach remains challenging.

Figure 2: NLP performance metrics



5. Results: Use Case

Several key areas of focus emerged as a result of the consultancy discussion. First, there is a clear need for a large, easily accessible corpus of free-text chief complaints that can form a standardised testbed for negation detection algorithm development and evaluation. Annotated data, in this context, consists of chief complaints annotated for *concepts* (e.g. vomiting, pain in chest) and the negation status of those concepts. See Figure 1 for a sample of chief complaints and their associated negated concepts (note that **(-)** indicates a negated term, **(+)** indicates a term that is not negated, and **(?)** indicates an uncertain term).

It is important that the annotation include both annotated clinical concepts and negation status to allow for the uniform evaluation and performance comparison of candidate negation detection

algorithms. Further, the annotated corpus should consist of several thousand (as opposed to several hundred) distinct and representative chief complaints in order to compare algorithms on a sufficient variety and volume of negation patterns. Corpus volume and variety are important due to the "long tail" problem. That is, most negation patterns are of a predictable format (e.g. "*denies abdominal pain*", "*no evidence of fever"*) yet there is a long tail of more unusual negation styles (e.g. "*bowel obstruction - resolved"*). While a corpus including *all* common and uncommon negation types is not possible, making inroads into the long tail is likely to improve performance for both rule-based and machine learning-based classifiers.

Unlike the general domain NLP field, which is increasingly focused on machine learning-based methods for language understanding, clinical NLP has relied on rule-based systems (e.g. NegEx & ConText). This is for two reasons. First, the annotation process is typically very expensive, especially for clinical data where medical expertise is required for annotation tasks, limiting the volume of annotated data generated. Machine learning approaches require a considerable volume of annotated data to work well, with the additional requirement that a substantial proportion of annotated data be excluded from training and "held out" for evaluation purposes. For situations in which annotated data are limited, rule-based systems can achieve better results. Second, machine learning algorithms are typically "black boxes" (i.e. the reasoning behind the classification decision is not interpretable by humans). In the medical and public health context, retaining interpretability is extremely important, particularly for decisions related to life-or-death and to resource allocation (money, time, and effort from trained personnel).

In addition to the need for annotated data-sets and appropriate NLP algorithms, a framework for evaluation is required. Typically, NLP algorithms are evaluated using *precision* (also known as *positive predictive value*), *recall* (also known as *sensitivity*), and *F-score* the harmonic mean of precision and recall. See Figure 2 for an explanation of these metrics.

Evidence currently suggests that shallow rule-based methods (i.e. methods that do not utilise deep syntactic and semantic parsing) in the vein of NegEx remain competitive for clinical negation detection [27].

5.1. Shallow supervised machine learning method for negation detection

Unlike rule-based approaches that encode knowledge with specific observations such as IF ... THEN rules, the supervised machine learning approaches depends heavily on the existence of training data. Training data is required in order that the algorithm can "learn" from provided examples to build classifiers that predict unseen inputs. Machine learning approaches consist of three major components: (1) training dataset, (2) at least one supervised machine learning algorithm, e.g., Support Vector Machine, Conditional Random Field, and (3) feature sets which serve as basic variables with which an algorithm builds a classifier. Each component contributes substantially to the quality of classifiers. Negation detection in clinical text using machine learning has been studied and reported promising results for some data sets [31].

The performance of machine learning approaches is dependent on the data sets used for training and evaluation. It has been shown that a classifier can perform well when trained and tested in the same environments or hospital, but perform poorly when applied to different hospitals [32]. This poor generalisability of algorithm performance is due to variability in practices across institutions and geographical areas. Compared to rule-based systems for negation detection, the machine learning approaches have limitations in both generalisability and customisability, though machine learning approaches have outperformed rule-based systems when training data are sufficient and features are appropriate [33].

With the emergence of *deep learning* (i.e. neural network-based machine learning algorithms with multiple layers), there are now several studies that have focused on using this approach for negation detection in general English [31], although studies in the clinical domain are less straightforward due to this methodology's need for substantial amounts of training data and computing resources.

5.2. Ensemble Methods

For negation detection, coupling machine learning with lexical and syntactic feature sets in an ensemble manner has proven at least as accurate as lexically-based approaches alone. This finding has been demonstrated by several machine learning-based NLP systems as part of the 2010 i2b2/VA challenge on identifying concepts, assertions, and relations in clinical text for accurately identifying negated concepts [34]. The 2010 i2B2/VA challenge dataset includes two corpora: the Beth Israel Deaconess Medical Center (BIDMC) Corpus of discharge summaries and the Partners Healthcare (PH) Challenge Corpus of discharge summaries, for which several ensemble classifiers have been benchmarked. For example, the Statistical Assertion Classifier (StAC) is a support vector machine trained with lexical and syntactic features to classify entities for both negation and uncertainty for two clinical corpora: the aforementioned 2010 i2b2/VA challenge corpora and the Computational Medicine Center (CMC) Corpus of radiology notes [35]. When compared to an extended version of NegEx called ENegEx, Uzuner et al. [35] observed that StAC outperformed ENegEx using a ± 4 word window and section headings. From their experiments, they also concluded that StAC can be applied without modification to new corpora, achieving similar performance to ENegEx.

In a complementary study of Mitre's CARAFE [36], an ensemble approach of Conditional Random Fields, Maximum Entropy, and manually crafted rules was trained and applied to the 2010 i2b2/VA challenge dataset. Clark et al. [36] observed that a baseline approach of word features derived from the concept to be classified with word features within a ± 3 token window of the concept of interest accounted for most of the accuracy achieved (91% F1-measure). However, they improved CARAFE's accuracy (93% F-measure, 93% precision, 93% recall) leveraging additional rich linguistic features: document structure, sentential structure, and semantic attributes of words in the sentence similar to those implemented by NegEx.

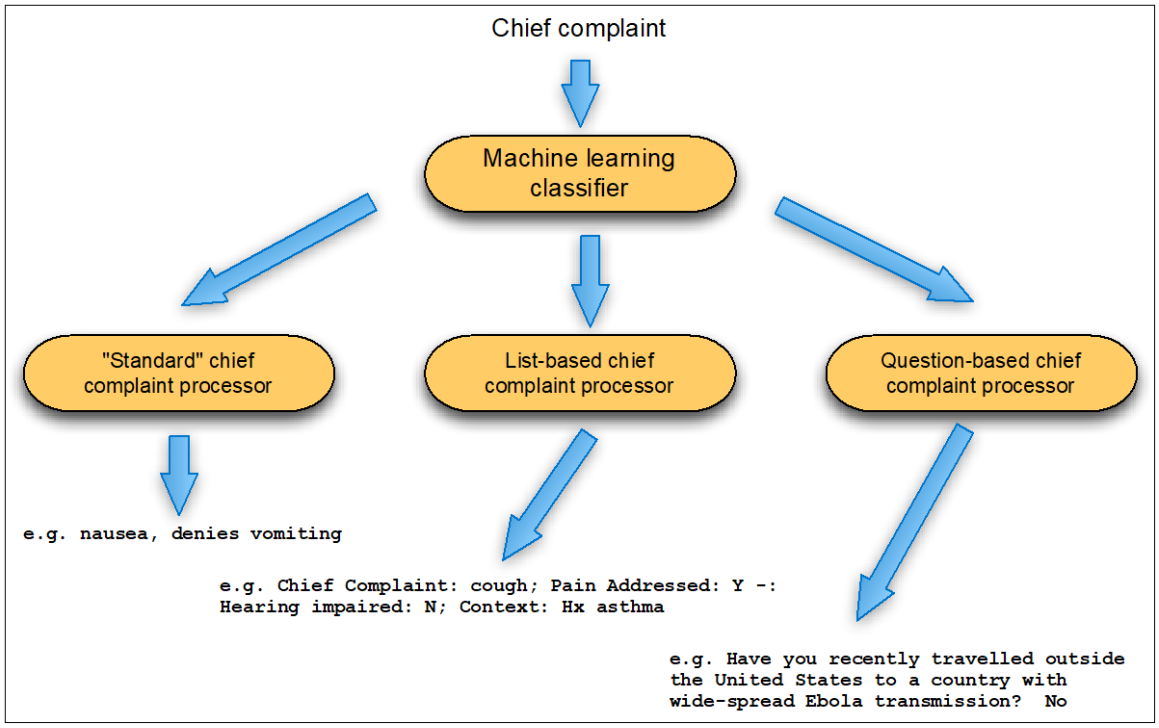
Performance improvement from the ensemble approach has been observed in other studies. For instance, NegEx Features Kernel (NF) is an ensemble approach that leverages feature outputs and rules from the NegEx algorithm to inform a linear kernel function using LibSVM (an implementation

of the Support Vector Machine algorithm) to support negation detection. Tested on the 2010 i2B2/VA challenge dataset, Shivade et al. [28] observed that NF performed with improved recall (90% precision; 90% recall) over NegEx (90% precision; 80% recall), with additional improvements when NF is coupled with a bag-of-words kernel and semi-supervised models.

5.3. Hybrid Methods

Given the heterogeneity of chief complaint data (see Figure 3) it is likely that different approaches will be required for different chief complaint types. For example, standard chief complaints (e.g. *vomiting, denies headache*), list-templated chief complaints (e.g. *nausea:n;vomiting;n*), and question-based chief complaints will require different approaches. We suggest the use of an initial filtering machine learning classifier to identify chief complaint type (e.g. the TagLine system [37]), followed by routing to the appropriate NLP algorithm.

Figure 3: General Purpose ML classifier



6. Discussion

Participants in the consultancy reported that the consultancy was effective in communicating the growing need for automated negation detection for syndromic surveillance with free-text chief complaints and other, more complex text fields. Conversely, public health practitioners appreciated summary explanations of deep and shallow, pure and hybrid algorithmic approaches potentially

applicable to their datasets. The opportunity for these practitioners and for solution developers from both academia — primarily computer scientists and biomedical informatics researchers — and industry was judged valuable by both participant types. Inclusion of representatives from several local and state health departments helped participants develop a sense of the various different negation detection and text processing challenges and practical approaches utilised across jurisdictions. In fact, some departments had avoided processing triage notes and other more complex fields for lack of trusted algorithms.

The consultancy had a number of limitations. First, a relatively short duration (1.5 days), in conjunction with an ambitious agenda, led to a sense among participants that some topics remained under-explored. For example, some participants suggested that machine learning approaches to negation detection could have been examined in more depth during the meeting. Second, a major barrier to developing text processing methods is the limited availability of annotated (coded) datasets for training and evaluating algorithms. A limited amount of Utah Department of Health free-text chief complaint data was obtained for sharing with consultancy participants, but this dataset was not annotated due to issues related to cost and time, and hence the utility of the dataset for developing and testing solutions was limited. Data provided by the other health departments consisted of free-text sample sets that were illustrative but not sufficient for calculation of statistical performance measures.

7. Conclusion

The consultancy was stimulating and eye opening for both technology developer and public health practitioner attendees. Developers unfamiliar with the everyday health-monitoring context gained an appreciation of the difficulty of deriving useful indicators from chief complaints. Also highlighted was the challenge of processing triage notes and other free-text fields that are often unused for surveillance purposes. Practitioners were provided with concise explanations and evaluations of recent NLP approaches applicable to negation processing. The event afforded direct dialogue important for communication across professional cultures. Key challenges to achieving enhanced processing of free-text data fields for routine population health monitoring were presented and explored. Most prominent among these challenges was the lack of labelled, authentic surveillance datasets; validation and user acceptance of opaque machine-learning algorithms that may be required for high-accuracy classification of triage note and similar text fields; and inter-regional and even inter-facility differences in free-text conventions that may confound the portability of successful classifiers. The intent of the consultancy planners was that dialogue would promote greater understanding and direct further efforts towards the goal of improving syndromic surveillance systems.

Acknowledgements

The organisation, preconference calls, and the consultancy itself were supported and funded by the Defense Threat Reduction Agency. Contents of this report are solely the responsibility of the authors and do not necessarily represent the official view of the Defense Threat Reduction Agency.

SV's contribution was supported by the Swedish Research Council (2015-00359) and the Marie Sklodowska Curie Actions, Cofund, Project INCA 600398.

References

1. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 34(5), 301-10. PubMed https://doi.org/10.1006/jbin.2001.1029

2. Faigen Z, Deyneka L, Ising A, Neill D, Conway M, et al. 2015. Cross-disciplinary consultancy to bridge public health technical needs and analytic developers: asyndromic surveillance use case. *Online J Public Health Inform*. 7(3), e228. PubMed https://doi.org/10.5210/ojphi.v7i3.6354

3. Reid M, Gunn J, Shah S, Donovan M, Eggo R, et al. 2016. Cross-disciplinary consultancy to enhance predictions of asthma exacerbation risk in Boston. *Online J Public Health Inform*. 8(3), e199. PubMed https://doi.org/10.5210/ojphi.v8i3.6902

4. Burkom HS. 2017. Evolution of public health surveillance: status and recommendations. *Am J Public Health*. 107(6), 848-50. PubMed https://doi.org/10.2105/AJPH.2017.303801

5. Purtle J, Field RI, Hipper T, Nash-Arott J, Chernak E, et al. 2017. The impact of law on syndromic disease surveillance implementation. *J Public Health Manag Pract*. PubMed

6. Choi J, Cho Y, Shim E, Woo H. 2016. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health*. 16(1), 1238. PubMed https://doi.org/10.1186/s12889-016-3893-0

7. Conway M, Dowling JN, Chapman WW. 2013. Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America. *J Biomed Inform*. 46(4), 734-43. PubMed https://doi.org/10.1016/j.jbi.2013.04.003

8. Scholer MJ, Ghneim GS, Wu S, Westlake M, Travers DA, et al. 2007. Defining and applying a method for improving the sensitivity and specificity of an emergency department early event detection system. *AMIA Annu Symp Proc*. (Oct), 651-55. PubMed

9. Lawson BM, Fitzhugh EC, Hall SP, Franklin C, Hutwagner LC, et al. 2005. Multifaceted syndromic surveillance in a public health department using the early aberration reporting system. *J Public Health Manag Pract*. 11(4), 274-81. PubMed https://doi.org/10.1097/00124784-200507000-00003

10. Hiller KM, Stoneking L, Min A, Rhodes SM. 2013. Syndromic surveillance for influenza in the emergency department-A systematic review. *PLoS One*. 8(9), e73832. PubMed https://doi.org/10.1371/journal.pone.0073832

11. Travers DA, Haas SW. 2004. Evaluation of emergency medical text processor, a system for cleaning chief complaint text data. *Acad Emerg Med*. 11(11), 1170-76. PubMed https://doi.org/10.1197/j.aem.2004.08.012

12. UNC - School of Nursing. About EMT-P. Available from: http://nursing.unc.edu/research/emtp/.

13. Harkema H, Dowling JN, Thornblade T, Chapman WW. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*. 42(5), 839-51. PubMed https://doi.org/10.1016/j.jbi.2009.05.002

14. Chapman BE, Lee S, Kang HP, Chapman WW. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *J Biomed Inform*. 44(5), 728-37. PubMed https://doi.org/10.1016/j.jbi.2011.03.011

15. Skeppstedt M. 2011. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. *J Biomed Semantics*. 2(Suppl 3), S3. PubMed https://doi.org/10.1186/2041-1480-2-S3-S3

16. Afzal Z, Pons E, Kang N, Sturkenboom M, Schuemie M, et al. 2014. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC Bioinformatics*. 15(1), 373. PubMed https://doi.org/10.1186/s12859-014-0373-3

17. Deléger L, Grouin C. Detecting negation of medical problems in French clinical notes. In: Proc. of the 2nd ACM SIGHIT Symposium on International Health Informatics; 2012. p. 697–702.

18. Costumero R, Lopez F, Gonzalo-Martín C, Millan M, Menasalvas E. An approach to detect negation on medical documents in Spanish. In: Ślezak D, Tan AH, Peters JF, Schwabe L, editors. Brain Informatics and Health: International Conference, BIH 2014, Warsaw, Poland, August 11-14, 2014, Proceedings. Springer International Publishing; 2014. p. 366–375.

19. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. 2017. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. *J Am Med Inform Assoc*. 24(3), 607. PubMed

20. Sohn S, Wu S, Chute CG. 2012. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science*. 2012, 1-8. PubMed

21. Gkotsis G, Velupillai S, Oellrich A, Dean H, Liakata M, et al. Don't let notes be misunderstood: a negation detection method for assessing risk of suicide in mental health records. In:

Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. San Diego, CA, USA: Association for Computational Linguistics; 2016. p. 95–105. Available from: http://www.aclweb.org/anthology/W16-0310.

22. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, et al. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *J Biomed Inform*. 54, 213-19. PubMed https://doi.org/10.1016/j.jbi.2015.02.010

23. Mutalik PG, Deshpande A, Nadkarni PM. 2001. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J Am Med Inform Assoc*. 8(6), 598-609. PubMed https://doi.org/10.1136/jamia.2001.0080598

24. Huang Y, Lowe HJ. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *J Am Med Inform Assoc*. 14(3), 304-11. PubMed https://doi.org/10.1197/jamia.M2284

25. Tanushi H, Dalianis H, Duneld M, Kvist M, Skeppstedt M, et al. Negation scope delimitation in clinical text using three approaches: NegEx, PyConTextNLP and SynNeg. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) NEALT Proceedings Series 16. Oslo, Norway; 2013.

26. Uzuner Ö, South BR, Shen S, DuVall SL. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 18(5), 552-56. PubMed https://doi.org/10.1136/amiajnl-2011-000203

27. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, et al. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PLoS One*. 9(11), e112774. PubMed https://doi.org/10.1371/journal.pone.0112774

28. Shivade C, de Marneffe MC, Fosler-Lussier E, Lai AM. Extending NegEx with kernel methods for negation detection in clinical text. In: Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015). Denver, Colorado; 2015. p. 41–46.

29. Ou Y, Patrick J. 2015. Automatic negation detection in narrative pathology reports. *Artif Intell Med*. 64(1), 41-50. PubMed https://doi.org/10.1016/j.artmed.2015.03.001

30. Mukherjee P, Leroy G, Kauchak D, Rajanarayanan S, Diaz DYR, et al. 2017. NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. *J Biomed Inform*. 69, 55-62. PubMed https://doi.org/10.1016/j.jbi.2017.03.014

31. Fancellu F, Lopez A, Webber B. Neural networks for negation scope detection. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016. p. 495–504.

32. Ou Y, Patrick J. 2015. Automatic negation detection in narrative pathology reports. *Artif Intell Med*. 64(1), 41-50. PubMed https://doi.org/10.1016/j.artmed.2015.03.001

33. Agarwal S, Yu H. 2010. Biomedical negation scope detection with conditional random fields. *J Am Med Inform Assoc*. 17(6), 696-701. PubMed https://doi.org/10.1136/jamia.2010.003228

34. Uzuner Ö, South BR, Shen S, DuVall SL. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*. 18(5), 552-56. PubMed https://doi.org/10.1136/amiajnl-2011-000203

35. Uzuner Ö, Zhang X, Sibanda T. 2009. Machine learning and rule-based approaches to assertion classification. *J Am Med Inform Assoc*. 16(1), 109. PubMed https://doi.org/10.1197/jamia.M2950

36. Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B, et al. 2011. MITRE system for clinical assertion status classification. *J Am Med Inform Assoc*. 18(5), 563-67. PubMed https://doi.org/10.1136/amiajnl-2011-000164

37. Finch DK, McCart JA, Luther SL. 2014. TagLine: information extraction for semi-structured text in medical progress notes. *AMIA Annu Symp Proc*. 2014, 534-43. PubMed