Self-Supervised Learning for Medical Web Disease Reporting Events Detection

Avaré Stewart
L3S Research Center
Hannover, German
stewart@L3S.de

Wolfgang Nejdl
L3S Research Center
Hannover, German
nejdl@L3S.de

ABSTRACT

In Web Science, Social Media Based Epidemic Intelligence has emerged as a type of medical intelligence gathering that supports health officials in routinely identifying potential health threats from more dynamic sources of information. State-of-the-art supervised approaches for SM-EI suffer from the high costs associated with manually labeling training examples.

This paper addresses the aforementioned problem by building a *self-supervised classifier*, one that labels its own training examples. Our results show that a self-supervised classifier, which discriminately selects its support vectors at each iteration achieves an F-measure of 78%. These results are comparable with existing, state-of-the-art systems, which rely exclusively on labeled data to build a classifier.

Categories and Subject Descriptors

H.4.2 [**Information Systems Application**]: Types of Systems

General Terms

Algorithms,Experimentation

Keywords

Self-Supervised Transfer Learning, Epidemic Intelligence

1. INTRODUCTION

Social Media Based Epidemic Intelligence (SM-EI) has emerged as a type of medical intelligence gathering within Web Science that aims to support health officials in routinely identifying potential health threats from more dynamic sources of information, such social media. Since it is infeasible for any person to manually comb through text to search for facts related to disease reporting events, SM-EI systems rely upon some form of automation (classification) for detecting disease reporting events (DR-events) from information sources that are more timely, and have a lower publication barrier.

State-of-the-art, supervised approaches [5, 6] suffer from a high initial and sustainability costs since, many training examples are needed to build a good classifier and for dynamic content training data can become easily outdated [3].

2. PROPOSED APPROACH

We address the aforementioned problem by proposing a new approach for building a supervised classifier to detect disease reporting events in dynamic web sources. As depicted in Figure 1, instead of building a classifier strictly from manually labeled data, we exploit outbreak reports[1], to build a *self-supervised classifier* - one that labels its own training examples.

We tackle the problem for SM-EI by relying upon the data in one domain (outbreak reports) to build a self-supervised classifer (i.e.,*Model Building* and *Model Selection*), which can ultimately be deployed in another (blogs). The underlying intuition behind our approach is that the moderated sources of outbreak reports, acts as a type of "interlingua", which constrains the pattern a disease reporting sentence can have within a dynamic source.

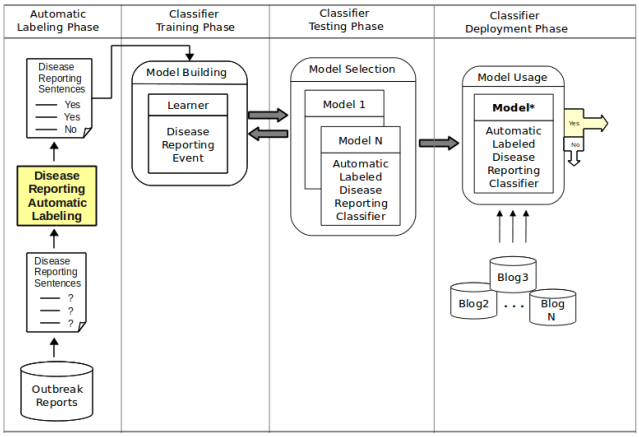

Figure 1: Overview of the self-supervised classifer approach to Disease Reporting Event Detection.

[1]An outbreak report is a moderated source that relies on experts to filter and extract information about health threats.

2.1 Proposed Sentence Weighting Schemes

The problem faced in our setting is that, in general, the outbreak reports are not labeled. To address this, we propose the several sentence weighting schemes. The weighting schemes take into account: the sentence position; structural, as well as non-structural features; and sentence semantics (i.e., presence of task specific named entities).

3. EXPERIMENTS

The goal of our experiments is the evaluate the performance of a self-supervised classifier when applied to blogs, using named entities (AutoNER) and structural features (POS and POSVEC) against a baseline, in which no structural features are present (VEC). In our experiments, both ProMED-mail and WHO outbreak reports were obtained via a web crawl. Each document was processed using the Standford Parser to split and parse each sentence. Named entities were extracted using LingPipe dictionary matching algorithm. Sentences having a length below 12 and above 200 character were eliminated, based on a frequency distributions of sentence lengths. To test the self-supervised classifier, we selected the AvianFluDiary, data set. Of the 5,328 manually labeled AvianFluDiary sentences, we used the 729 positive cases found, and balanced this by selecting 729 random negative cases.

3.1 Data Summary

A subset of the ProMED-Mail (and WHO) sentences were was constructed by using automatic labeling using automatic labeling with named entities. In total for ProMED-mail are 3452 positive and 2342 negative sentences were automatically labeled; for WHO 1412 positive and 473 negative sentences were automatically labeled.

SVM Classifier Tuning: The classifier used in our work was based on the open source implementation of SVM-TK by Moschitti [4]. The classification features used to build the classifier were the Penn-Tree Bank, parts-of-speech parse tree (POS), the term vector (VEC), and their combination (POSVEC). We trained the self-supervised classifer on outbreak reports and then tested the classifier on the 729 manual labeled positive and negative examples and report the fmeasure performance.

3.2 Results

Table 1: Performance fmeasure for ProMED-mail and WHO self-supervised classifer for the VEC, POS and POSVEC features with named entities (AutoNER).

Feature Type	ProMED-mail	WHO
VEC	78.12	67.43
POS	71.83	72.88
POSVEC	78.86	72.86

Per outbreak source, we notice that the ProMED-mail self-supervised classifer performances best using the POSVEC and VEC features. The best performing self-supervised classifer for the WHO reports is POS and POSVEC. The vector performance for self-supervised classifer using ProMED-mail can be explained by the fact that ProMED-mail data contains more terms. The additional terms with the VEC feature also contribute towards the performance boost for the ProMED-mail POSVEC. On the other hand, for WHO, we see that the self-supervised classifer which used the VEC feature actually performs rather poorly; and also does not significantly boost the performance of the POSVEC beyond that of the POS feature.

Comparison with the State-of-the-Art: We compare the the performance of theself-supervised classifer to reported results for the similar sentence-level classification tasks [6, 5]. Work has been done by Zhang [6] for classifying disease reporting sentences. In their work, an F1-measure value of 76% is reported. When taking named entities into account, we obtain a comparable F1-measures of 78% (ProMED-mail), in contrast to their work, we use a self-supervised classifier. In other works statistical approaches for classifying disease-reporting sentences have been carried out, [1, 2]. In all cases, manually labeled data to build their models. Alternatively, we seek to go beyond the human effort associated with building a training set, while striving for comparable results with current statistical state-of-the-art systems.

4. CONCLUSION

In this paper, we have laid the groundwork for a new approach towards Disease Reporting classifier for Event-Based Epidemic Intelligence. Instead of building a classifier strictly from manually labeled data, we instead exploit outbreak reports for training data. Since the outbreaks reports are not explicitly labeled, we introduce the use of a self-supervised classifier, one that labels its own training data. We find that our self-supervised classifer is able to achieve an fmeasure as high as 78% by using a parts-of-speech as features and named entities. Also, we have shown that our self-supervised classifier is comparable with existing state-of-the-art systems.

5. REFERENCES

[1] M. Conway, N. Collier, and S. Doan. Using hedges to enhance a disease outbreak report text mining system. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 142–143, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[2] M. Conway, S. Doan, A. Kawazoe, and N. Collier. Classifying disease outbreak reports using n-grams and semantic features. *International Journal of Medical Informatics*, 78(12):e47–e58, 2009.

[3] G. Lakshmanan and M. Oberhofer. Knowledge discovery in the blogosphere: Approaches and challenges. *IEEE Internet Computing*, 14:24–32, March 2010.

[4] A. Moschitti. Making tree kernels practical for natural language learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[5] M. Naughton, N. Stokes, and J. Carthy. Sentence-level event classification in unstructured texts. *Inf. Retr.*, 13(2):132–156, 2010.

[6] Y. Zhang. *Automatic Extraction of Outbreak Information from News*. PhD thesis, University of Illinois at Chicago, 2008.