EXPLOITING THE LANGUAGE OF MODERATED SOURCES FOR CROSS-CLASSIFICATION OF USER GENERATED CONTENT

Avaré Stewart and Wolfgang Nejdl
Forschungszentrum L3S, Appelstr 9a, Hannover, Germany

Keywords: Automatic labeling, Cross-classification, Medical intelligence gathering.

Abstract: Recent pandemics such as Swine Flu have caused concern for public health officials. Given the ever increasing pace at which infectious diseases can spread globally, officials must be prepared to react sooner and with greater epidemic intelligence gathering capabilities. However, state-of-the-art systems for Epidemic Intelligence have not kept the pace with the growing need for more robust public health event detection. Existing systems are limited in that they rely on template-driven approaches to extract information about public health events from human language text.

In this paper, we propose a new approach to support Epidemic Intelligence. We tackle the problem of detecting relevant information from unstructured text from a statistical pattern recognition viewpoint. In doing so, we also address the problems associated with the noisy and dynamic nature of blogs by exploiting the language in moderated sources, to train a classifier for detecting victim reporting sentences in blog social media. We refer to this as Cross-Classification. Our experiments show that without using manually labeled data, and with a simple set of features, we are able to achieve a precision as high as 88% and an accuracy of 77%, comparable with the state-of-the-art approaches for the same task.

1 INTRODUCTION

Many factors in today's changing society contribute towards the continuous emergence of infectious diseases. In response, Epidemic Intelligence (EI) has emerged as a type of intelligence gathering which aims to detect events of interest to the public health from unstructured text on the Web.

In a typical EI framework, disease reporting events (i.e., victim, location, time, disease) are extracted from raw text. The events are then aggregated to produce signals, which are intended to be an early warning against potential public health threats. Epidemiologists use them to assess risk, or corroborate and verify the information locally and with international agencies.

Although there are numerous EI systems in existence, they are limited in two major ways. First, these systems focus mainly on using news and outbreak reports as a source of information (Hartley et al., 2009). However, in order to effectively provide a warning as early as possible, diverse information sources are needed, such as those from just-in-time crisis information blogs[1]. Disproportionately, blogs and other types of social media have not been considered in intelligence gathering. Secondly, the algorithms used in these systems typically detect disease related activity by relying upon predefined templates, such as keywords or regular expression. The drawbacks of a template-based approach is that given the variety of natural language, many patterns may be required and enumerating all possible patterns is costly. Moreover, the results typically lead to a low recall for identifying relevant events.

The first steps toward overcoming this limitation is to view the Epidemic Intelligence task in a new light, by: 1) including more diverse sources, such as blogs, and 2) using statistical approaches (e.g., statistical pattern recognition), to detect information about public health events. However, the automatic extraction of information from blogs remains a challenging task, because blogs are both noisy and dynamic (Moens, 2009).

In this paper we address these twofold challenges first by relying upon *comparable text*. We say that comparable text is one in which overlapping topics

[1] http://www.ushahidi.com

Stewart A. and Nejdl W..
EXPLOITING THE LANGUAGE OF MODERATED SOURCES FOR CROSS-CLASSIFICATION OF USER GENERATED CONTENT.
DOI: 10.5220/0003298105710576
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 571-576
ISBN: 978-989-8425-51-5

are discussed in a similar way. For such text, we assume that the languages used have common parts, and therefore, the linguistic structures in one corpus, can be identified in the other. Second, we exploit this notion by building a binary classifier, that has been trained from the victim-reporting sentences of one corpora, in our case outbreak reports. The training data is gathered through weak labeling, i.e., the training set is automatically built. A classifier trained on this training set is then used to detect disease-reporting sentences in blogs. We refer to this approach as Cross-Classification. The contributions of this work are: 1) an introduction of a Cross-Classification Framework for Epidemic Intelligence, and 2) an exploitation of outbreak reports for weak labeling.

The rest of the paper is structured as follows: the Cross-Classification approach is described in Section 2 and an evaluation is given in Section 3. In Section 4, related work is presented and in Section 5, the paper concludes with a discussion on future work.

2 CROSS-CLASSIFICATION FRAMEWORK

In this section, we describe our Cross-Classification approach and outline how the text of outbreak reports is exploited for weak labeling of blog post sentences. We also outline the properties that impact the quality of weak labeled sentences when using such sources.

2.1 Cross-classification

The Cross-Classification process is depicted in Figure 1. The comparable texts are two distinct corpora. We define one as the auxiliary source (outbreak reports) and the other one as the target source (blogs). Additionally, the figure shows the traditional way of classification in which the sentences of the target domain are labeled manually and used as training data (Direct Classification).

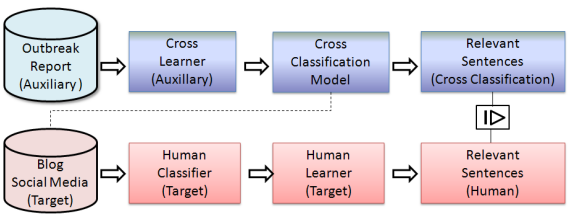

Figure 1: Overview of Cross-Classification.

An outbreak report is a moderated source that typically relies on experts to filter and extract information about health threats. Based on the underlying properties of the auxiliary data (see Section 2.2), we automatically label the sentences in the auxiliary data as positive or negative and use them for training a binary classifier. The resulting model is then applied on the target corpus in order to classify sentences.

The type of classifier we use is a tree-kernel based support vector machine (SVM). The reason for this choice is that for natural language tasks similar to ours, linguistic representations have proven to be successful for detecting disease reporting information (Zhang, 2008; Conway et al., 2009). One reason is that kernel methods are capable of generating a high number of syntactic features, from which the learning algorithm can select those most relevant for a specific application (Moschitti, 2006). This can help overcome the feature engineering needed with linguistic representations.

2.2 Weak Labeling for Training Data

Since training data is unavailable for our classification task, we apply weak labeling for gathering training material. In particular, the Sentence Position is exploited for selecting positive and negative examples. As in any classification task, the quality of the classifier is highly dependent on the training data and its noise. We also investigate further properties that may impact the quality of weak labeling, these include: Sentence Length, and Sentence Semantics.

Sentence Position. The position of information in text has been widely exploited in document summarization for news where the first sentences in an article or paragraph summarize the most important information (Lam-Adesina and Jones, 2001). Based on this, we model the auxiliary corpus as a sentence database, where each document, in the corpus is represented as an ordered sequence of one or more sentences. We adopt an approach to automatically label the sentences in each document, where the TopN sentences in a document are taken as positive cases, for a threshold value of N. Further we hypothesize that sentences appearing towards the end of the sequence, are less relevant, so the BottomN are automatically labeled as negative examples.

Sentence Length. The sentences in the auxiliary corpus, vary greatly in length, due to conjunction and phrases. Previous work using tree representations for sentences, has shown that longer sentences may contain too many irrelevant features, and over-fitting may occur, thereby decreasing the classification accuracy. In this light, we propose that sentence length is also an important aspect of Cross-Classification and inve-

stigate its impact on the quality of Cross-Classification in our experiments.

Sentence Semantics. Finally, we are interested in identifying victim-reporting sentences, where the definition used for victim reporting is based on the template for MedISys Disease Incidents[2]. This template includes disease, time, location, case, and the status of victims that have been extracted from the full text of news articles. We say a sentence is a victim-reporting one, if it contains a medical condition in conjunction with a victim, time, or location, where the case and status of a victim may be inferred from the context. The semantic information is thus represented by the presence of named entities (NEs) in the sentence. In our experiments, we also investigate if the presence of NEs is an important factor in choosing positive examples.

3 EXPERIMENTS

The goal of our experiments is to measure how well a Cross-Classifier can detect victim-reporting sentences within a blog. We train the classifier with three different feature sets and compare the results. Further, we experiment with three weak labeling properties for the outbreak reports, namely: Sentence Position, Sentence Length, and Sentence Semantics. We compare the performance of the Cross-Classifier to that of a traditional classifier, and state-of-the-art performance measures reported for a similar task, which use a considerable number of features and rely exclusively upon labeled data for training.

3.1 Experimental Setting

As target data, we selected the AvianFluDiary[3], a well known source within the "flu bloggers" community. The data was collected for a one year period: January 1 - December 31, 2009. For the auxiliary data, we used ProMED-mail[4], a global electronic reporting system, listing outbreaks of infectious diseases. This data was collected over a period of eight years: January 1, 2002 - December 31, 2009. Early experiments using the data from a single year, as the auxiliary data, showed poor results, due to the fact that there were too few documents (and hence sentences) to support weak labeling. In total, we collected 4,249 documents for AvianFluDiary and 14,665 for ProMED-mail.

The data for both the auxiliary and target domains was processed using the Standford Parser[5] to split and parse each sentence. The total number of sentences for AvianFluDiary was 44,723 and ProMED-mail 347,822. Sentences were further processed using OpenCalais[6] to extract NEs. In total, 3,300 documents and 34,752 sentences from AvianFluDiary and 10,026 documents and 127,314 sentences from ProMED-mail contained entities recognizable by OpenCalais.

Weak labeling for the auxiliary data was constructed using the Top5 sentences as positive examples. Roughly 25,000 sentences were included in the Top1, and the amount of training data increased by 25,000 sentences, as N increased. Further, we notice in our corpus that the bottom sentences tend to refer to additional web sites, so any bottom N sentences containing URLs where eliminated.

SVM Classification. The Cross-Classifier was based on SVM-TK (Moschitti, 2006), and the classification features used to build the classifier were the parts-of-speech parse tree (POS), the term vector (VEC), and their combination (POSVEC). Experiments using the vector space feature alone performed consistently below the POS and POSVEC, thus, we do not report them further in this work. Five random sets, from each TopN set were selected to train a classifier, and the results obtained for each classifier were averaged over these five trials.

Baseline. As a baseline, we compare the Cross-Classification approach to the traditional (or Direct Classification) method (see Figure 1), in which both the training and testing is done on manually labeled sentences of AvianFluDiary. Of the 5,328 sentences manually labeled, 729 were positive cases, showing a relatively low percentage of information-bearing sentences within the blog. The direct classifier was trained with equal amounts of positive and negative cases using a 10-fold cross-validation. The evaluation measures used were precision (P), recall (R), f1-measure (F) and accuracy (A), reported using a scale of 0 to 100%.

We next present our experiments: first we evaluate the Cross-Classification performance with respect to the three weak labeling properties of outbreak reports; then, the classification features are examined.

[2] http://medusa.jrc.it/medisys/helsinkiedition/all/home.html

[3] http://afludiary.blogspot.com/

[4] http://www.promedmail.org

[5] http://nlp.stanford.edu

[6] http://www.opencalais.com

Table 1: Cross-Classification performance for each of the topN (N=1···5) sentences using training sizes of 1K, 2K and 3K, and the POSVEC feature. Each of the topN entries shown is the result of averaging over 5 trials.

Size	N	P	R	F	A
1K	1	81.27	44.55	57.54	67.15
	2	79.71	65.51	71.85	74.39
	3	78.44	70.34	74.15	75.50
	4	77.90	75.75	***76.77***	***77.13***
	5	75.90	75.56	75.70	75.78
2K	1	***82.41***	48.34	60.93	69.01
	2	80.54	67.30	73.29	75.50
	3	78.80	73.17	75.85	76.73
	4	76.33	75.58	75.93	76.06
	5	76.09	75.50	75.79	75.88
3K	1	82.15	47.98	60.57	68.78
	2	79.49	67.63	73.06	75.07
	3	77.99	74.29	76.08	76.65
	4	76.57	76.16	76.35	76.42
	5	76.89	***76.30***	76.57	76.67

Table 2: Cross-Classification performance for each of the topN (N=1···5) sentences using training sizes of 1K, 2K and 3K, and the POS feature. Each of the topN entries shown is the result of averaging over 5 trials.

Size	N	P	R	F	A
1K	1	76.61	43.04	55.10	64.95
	2	76.72	64.25	69.86	72.33
	3	76.24	68.34	72.07	73.51
	4	75.68	73.50	74.54	74.94
	5	74.44	72.98	73.67	73.92
2K	1	***77.78***	46.45	58.16	66.60
	2	76.96	65.40	70.66	72.88
	3	76.57	71.00	73.67	74.62
	4	74.40	73.53	73.93	74.10
	5	75.08	73.83	74.44	74.66
3K	1	77.71	46.20	57.93	66.48
	2	76.14	65.62	70.46	72.51
	3	75.70	71.82	73.70	74.38
	4	74.64	73.52	74.05	74.25
	5	76.27	***74.07***	***75.14***	***75.49***

3.2 Weak Labeling Properties

3.2.1 Sentence Position

We evaluate how the position of the sentence within the document affects the performance of a weak labeler when compared against a random selection of sentences for a direct classifier. Tables 1, and 2 summarize the Cross-Classification performance using the features: POSVEC, and POS, when varying the training size (Size) from 1,000 (1K) to 3,000 (3K) sentences; using a fixed sentence length of 5 to 199 characters. The bold font in each table shows the maximum values obtained for each measure. The results clearly show that we obtain very good results without manual labeling - precision reaching 82.41% and recall 81.65%. Also, increasing the training size improves the results, because the classifier has more examples from which it can learn. Also, in terms of a performance trade-off, the Top1 and Top2 positions prove not to be the best, but instead Top3 or Top4 show better trade-off, as the precision becomes equal to the recall.

Random Selection. The results for the Cross-Classifier built from a random selection of sentences as training set, using the POSVEC feature is presented in Table 3. We notice that the random classifier performs significantly poorer than the one in which the weak labeling is used. This clearly suggests that the sentence order is a useful, yet simple criteria for weak labeling.

Table 3: Cross-Classification performance using a random selection of sentences and training sizes of 1K, 2K and 3K, for the POSVEC feature.

Size	P	R	F	A
1K	41.37	40.11	40.63	41.61
2K	***42.85***	***43.62***	***43.11***	***42.88***
3K	33.37	32.66	32.94	33.40

Table 4: Direct Classification performance for the POSVEC feature.

P	R	F	A
86.50	90.42	88.32	88.06

Direct Classification. In the Direct Classification, we used the manually labeled data of the target corpus as a training set, and the results are shown in Table 4. When only the sentence position is taken into account, we see that the overall performance of the Direct Classifier is significantly better than the Cross-Classifier in terms of its recall and f1-measure. Yet, in terms of precision, the Cross-Classifier obtains as much as 82.41% (see Table 1), in comparison with the Direct Classifier, which is 86.50%. Also, the recall is much higher for the Direct Classifier. This is to be expected, given the errors inherent in weak labeling.

Finally, we compare the Cross-Classification to reported results for the same task (Zhang, 2008) where the highest f1-measure value obtained is just above 76%. When considering sentence position alone, we obtain comparable f1-measures of 76.77% (see Table 1).

3.2.2 Sentence Length

In order to determine if longer sentences impact the performance of the weak labeled classifier, we created a second partition of data based on sentence lengths in the interval of [200...1024]. Although the range of this interval is quite large in comparison with the interval [5...199], the actual number of sentences is much smaller. Interval [5...199] contains 96,851 sentences in the Top5 set, whereas interval [200...1024] contains 19,368. Figure 2 shows the average over the Top5 results for the POSVEC feature.

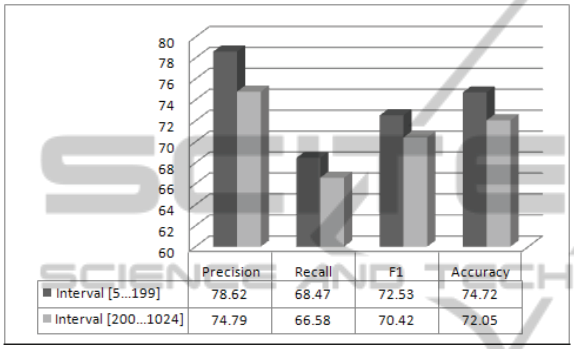

Figure 2: Cross-Classification performance for sentence lengths in the ranges of [5...199] and [200...1024] using the POSVEC feature.

It can be seen, that using longer sentences results in a classifier with lower overall performance. We believe this to be the case because more noise is introduced as longer sentences include clauses and parenthetical information, which are not directly related to victim reporting.

3.2.3 Sentence Semantics

We evaluate the performance of the Cross-Classifier in the presence of selected NEs, which are relevant for victim-reporting. These entity types include: location and medical condition. In order to make as much distinction as possible between the positive and negative NE-examples, only the TopN sentences containing the named entities were chosen for training, whereas the BottomN sentences *not* containing those entities were used. In Figure 3, the Cross-Classification results are obtained using sentence lengths in the interval of [5...199] characters, averaging over the Top5 results for the POSVEC feature and using a training size of 3K. We notice that filtering weak labeled sentences with respect to the presence of NEs yields a significantly higher precision when compared with no NEs. Thus, NEs are useful for filtering noise that is present in the weak labeling examples.

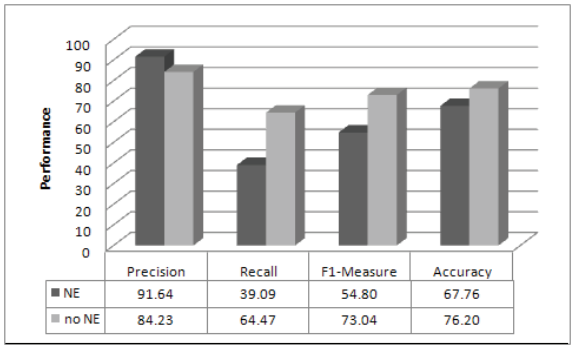

Figure 3: Cross-Classification performance with Named Entities (NEs), and without Named Entities (no NEs). The Top5 (N=5) sentences were used, and averaged over 5 trials, for a training size of 3K, with the POSVEC feature.

3.3 Discussion

The experiments presented above allow us to see that Sentence Position, Sentence Length and Sentence Semantics do, in fact, impact the ability of a weak labeling Cross-Classifier to detect the relevant victim-reporting sentences and several points should be noted regarding each.

- **Sentence Length.** Although we have experimented with two different ranges for the sentence lengths, a closer examination should be made to determine the minimum and maximum lengths that optimize the precision and recall. Even so, we already can draw the conclusion that smaller sentences (range [5...199]) already bring the benefit of being able to distinguish between the information bearing sentences, shown by our results.
- **Sentence Position.** It should be noted that the Sentence Position is not independent of Sentence Length. As mentioned, the shorter sentences that appear in the Top1 often consist of titles or even concise summaries of the article. Refinements which optimize the length of the sentence, should also take this into account.

4 RELATED WORK

In the area of Epidemic Intelligence, approaches for classifying disease-reporting sentences have been carried out, where a number of features are used and different types of techniques, such as conditional random fields and Naive Bayes networks (Zhang, 2008; Conway et al., 2009). In all cases, the authors use manually labeled data to build their models. In our work, we seek to go beyond the human effort associa-

ted with building a training set for blogs and social media, while striving for comparable results with these state-of-the-art systems.

Transfer Learning. Transfer Learning allows the domains, tasks and distributions for a classifier's training and test data to be different (Pan and Yang, 2009). The sub-area of transfer learning most similar to our work is transductive transfer learning, where neither the source nor target data is labeled. In this case, methods are sought to first automatically label the source data.

Automatic Labeling. Work has been done in several areas (Tomasic et al., 2007; Fuxman et al., 2009) to reduce the human labeling effort; where automatic Labeling has been achieved with **weak labeling**. In one such work, (Tomasic et al., 2007) wild labels (obtained from observing users) provide the basis for generating weak labels. Similar to our work, weak labels are distinguished from gold labels, which are generated by a human expert. The weakly-labeled corpus is used to train machine-learning algorithms that are capable of predicting the sequence and parameter values for the actions a user will take on a new request. In other work automatic labeling is accomplished by first defining a set of criteria a potential corpora must have in order to support the automatic labeling process (Fuxman et al., 2009). To date, none of the work based on automatic labelling or a transfer learning approach, consider the task of Epidemic Intelligence.

5 CONCLUSIONS AND FUTURE WORK

In this paper we have demonstrated that with our *Cross-Classification* framework, it is possible to use *comparable text*, such as outbreak reports as automatically labeled data for training a classifier that is capable of detecting the victim-reporting sentences in a blog.

As with any automated labeling process, the examples are subject to noise and error. We investigated how this noise can be reduced and evaluated the quality of such weak labeled sentences using three properties: Sentence Position, Sentence Length and Sentence Semantics. With no effort in human labeling and minimalistic feature engineering, we were able to build a Cross-Classifier, which achieved a precision as high as 88%. The impact of this work is that the noisy sentences in blogs, and possibly other types of social media, can be appropriately filtered to support epidemic investigation.

Cross-Classification has shown to be promising for data in which the topic is rather focused. As a future work, we will apply the approach to more diverse and topic-drifting blog posts. Further, we seek to generalize the results presented here, describing the conditions under which corpora can be considered comparable. This would help in automatically selecting the appropriate auxiliary and target corpora for Cross-Classification. In this work, we have assumed the presence of a high quality data set that lends itself to weak labeling. As further work, we plan to consider cases in which a Cross-Classifier can be built from less volume of data, for example, by using a bootstrapping approach.

REFERENCES

Conway, M., Collier, N., and Doan, S. (2009). 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. Association for Computational Linguistics.

Fuxman, A., Kannan, A., Goldberg, A. B., Agrawal, R., Tsaparas, P., and Shafer, J. (2009). Improving classification accuracy using automatically extracted training data. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1145–1154, New York, NY, USA. ACM.

Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., Thinus, G., and Lightfoot, N. (2009). The landscape of international event-based biosurveillance. *Emerging Health Threats*.

Lam-Adesina, A. M. and Jones, G. J. F. (2001). Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 1–9, New York, NY, USA. ACM.

Moens, M.-F. (2009). Information extraction from blogs. In Jansen, B. J., Spink, A., and Taksa, I., editors, *Handbook of Research on Web Log Analysis*, pages 469–487. IGI Global.

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

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 99.

Tomasic, A., Simmons, I., and Zimmerman, J. (2007). Learning information intent via observation. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 51–60, New York, NY, USA. ACM.

Zhang, Y. (2008). *Automatic Extraction of Outbreak Information from News*. PhD thesis, University of Illinois at Chicago.