Anomaly Detection in Text: The Value of Domain Knowledge

Raksha Kumaraswamy
Indiana University, Bloomington
rakkumar@indiana.edu

Anurag Wazalwar
Indiana University, Bloomington
anurwaza@indiana.edu

Tushar Khot
Allen Institute for Artificial Intelligence
tushar@cs.wisc.edu

Jude Shavlik
University of Wisconsin-Madison
shavlik@cs.wisc.edu

Sriraam Natarajan
Indiana University, Bloomington
natarasr@indiana.edu

Abstract

We consider the problem of detecting anomalies from text data. Our hypothesis is that as with classical anomaly detection algorithms, domain-specific features are more important than the linguistic features. We employ the use of first-order logic and demonstrate the effectiveness of useful domain knowledge in two domains. Our results show that the domain-specific features are more predictive and that the relational learning methods exhibit superior performance.

Introduction

Anomaly detection has been defined as the problem of finding patterns in the data that do not conform to the expected (normal) behavior (Chandola, Banerjee, and Kumar 2009). This problem has important applications from traffic patterns (Barria and Thajchayapong 2011) to security (Zhang and Zulkernine 2006).We consider a supervised approach to identifying anomalies in text. Our definition of anomaly also follows the standard definition of "deviating from normal (expected) situations". We are interested in document classification, i.e., identifying documents that deviate from the normal (Guthrie 2008). Previous approaches to anomaly detection from text mostly constructed lexical features and employed a classifier (Manevitz and Yousef 2007).

We hypothesize that the definition of anomaly in the context of textual data depends on the domain of interest. For instance, when reading sports articles, an example anomaly is when a low-ranked team playing away from home defeats a top-ranked team. This "upset" can be identified based on the knowledge of the teams, their relative rankings etc. and not necessarily on the lexical features. Similarly, in the recent unfortunate incident of the missing flight, the size of the aircraft and the zone in which it was flying are crucial to identifying it as an anomaly. When tagging anomalies in a question forum, the context of a question (for example, requesting solutions to a homework problem) is crucial. The common theme across all these situations is that specific information about the domain is more important than the lexical features. Such domain knowledge can be naturally provided in first-order logic (FOL) as features or advice rules. Consequently, it is easier to learn using richer representations such as Statistical Relational Learning (SRL) (Getoor and Taskar 2007). We employ a recently successful learning algorithm called relational functional gradient-boosting (RFGB) (Natarajan et al. 2012) for learning to predict the anomalies and show that it outperforms standard approaches.

We make a few key contributions in this work: (1) we show that using domain knowledge can substantially improve the detection of anomalies in text and (2) we also show that simply looking at syntactic features can greatly reduce the predictive performance of automated anomaly detection (3) finally, we evaluate using two domains - a literature domain, inspired by the work of Guthrie (2008), where the goal is to identify text that does not belong to a particular author, and a flight domain where the goal is to read about flight incidents and identify the relatively "unexpected" incidents.

Background and Related Work

Anomaly detection

Chandola et al (2009) provide an overview of the standard anomaly detection methods across multiple domains, however we focus only on textual domains. Guthrie's work (2008) on unsupervised anomaly detection from text proposes the use of primarily stylistic features of authors to identify anomalous documents or excerpts. This approach works best on literary works as it focuses on the syntax of the text, reading difficulty etc. This work was extended by Mahapatra et al (2012) whose method exploits contextual information from external sources and uses this as a post-processing step to improve or correct anomaly predictions. Manevitz & Yousef (2002) have demonstrated the use of both one-class SVM classifiers as well as neural networks, using only positive examples, to detect outliers in text. Another interesting method by Baker et al (1999) proposes a heirarchical probabilistic model for novelty detection in text. All these methods use a propositional representation and employ standard learning algorithms. We instead use a SRL method for this task.

Relational Functional-Gradient Boosting

We now present details of the learning algorithm we employ in this work called relational functional-gradient boosting

(RFGB) (Natarajan et al. 2012). Let $\mathbf{x}$ denote the features which in our case are lexical and syntactic features and y the target relations (example, anomaly(document)). Note that these are FOL predicates but we use variables for notational simplicity. The goal is to fit a model $P(y|\mathbf{x}) \propto e^{\psi(y,\mathbf{x})}$ for every target relation y. Functional gradient ascent starts with an initial potential ψ_0 and iteratively adds gradients Δ_m (i.e., instead of working in the parameter space, it performs gradient-descent in the function space). These gradients over the potential function are approximated by computing the gradients for each training example, i.e.,

$$\frac{\partial \log P(y_i; \mathbf{x_i})}{\partial \psi(y_i = 1; \mathbf{x_i})} = I(y_i = 1; \mathbf{x_i}) - P(y_i = 1; \mathbf{x_i}) \qquad (1)$$

Intuitively, this is the difference between the true value and the predicted probability of the current model (Dietterich, Ashenfelter, and Bulatov 2004). This set of local gradients reweighs the set of training examples and a new relational regression tree (RRT) (Blockeel and Raedt 1998) is fitted to these examples at each step. The final model is a weighted combination of all the RRTs. To learn the model for a target relation, say $anomaly(doc)$, we start with an initial model ψ_0 [1]. Then, we learn a RRT to fit the regression examples and add it to the current model. We now compute the gradients based on the updated model and repeat the process. In every subsequent iteration, we *fix* the errors made by the model. One of the key advantages of RFGB is that we learn a *large number of short RRTs*(each RRT can be considered to be capturing a set of rules).

Domain-dependent Anomaly Detection

To validate our hypothesis, we designed two types of domain knowledge that can be exploited by the anomaly detection system. The first kind is the standard learning approach of designing "good" domain dependent features using predicate logic. The second kind is "advice" where the system is provided with domain-specific background knowledge. Inspired by several successes in NLP tasks, we designed clauses for a Markov Logic Network (MLN) (Domingos and Lowd 2009) and used the resulting MLN for anomaly detection. We designed two domains for evaluation of RFGB and MLN based SRL approaches against standard ML methods.

Flight Domain: Our first domain is the identification of *anomalous flight incidents* from text. We created a dataset consisting of 45 news articles reporting different flight incidents. Anomalies in this domain refer to unexpected flight incidents such as missing or crashed passenger aircrafts with more than 100 passengers in non-war zones. Of the 45 articles, we identified 18 articles as anomalous.

To identify flight incidents from text, we modified the parser to check for presence of certain words which are homologous to the word 'incident' in the articles, and used them as features for the article. We also extracted the number of passengers in the flight, if mentioned in the article, and used it as a feature. In addition, we also included unigram and bigram features, i.e., tfIdf scores, and word presence, along with features corresponding to parse trees and dependency graphs, encoded in FOL predicates. These parsing based features are domain knowledge features of type 1 - "good" domain dependent features represented in predicate logic. To incorporate domain knowledge features of type 2 - "advice", we decided to include existing knowledge about different flights from free external knowledge repositories like Freebase and Wikipedia. Since the capacity of a flight may not be always mentioned in an article, we obtained a list of aircraft models, manufacturers and their capacities (source: Freebase). If an aircraft model is mentioned in the article then its number of passengers is assumed to be its capacity as obtained from these "advice" predicates. We also obtained a similar list of known warzones (source: Wikipedia) and added these as "advice". A sample of the features used by our system have been described in Table 1.

Predicate	Explanation
incidentWord(*doc*, *inc*)	*inc* is the incident that occurs in document *doc*
takeOffLocation(*doc*, *loc*)	*loc* is the location from which the flight took-off
incidentLocation(*doc*, *loc*)	*loc* is the location at which the flight incident occurs
warzone(*loc*)	indicates that *loc* is a warzone
aircraftType(*doc*, *type*)	*type* is the type of the aircraft in the document *doc*

Table 1: Features for the Flight Domain.

Anomalous Excerpts Domain: Inspired by Guthrie's work (2008) on anomaly detection in a literature domain, we created a dataset with anomalous literary excerpts. We selected excerpts with $20 - 30$ sentences from Sir Arthur Conan Doyle's books to create the set of normal documents. To create anomalous documents, we introduced a sentence, at random, from Jane Austen's books in some excerpts from Doyle. As "good" domain knowledge features, we created standard unigram and bigram features for this domain, since the words used by the authors (especially names e.g., Holmes vs Emma/Catherine) could be very different and significant for detection of anomalies in this domain. To include "advice" in this domain we included standard NLP *readability measures*, as suggested by Guthrie (2008). Similar to Guthrie et al., we have used four such readability measures and present a sample in Table 2. These readability measures essentially capture the style of an author and since they are dependent on features specific to the author's writing style, they should assist in differentiating these authors efficiently, and hence serve as "advice" in only this domain. They are calculated using statistics computed from text such as number of words per sentence, number of polysyllables per sentence etc.

Experimental Results

The key questions that we aim to answer empirically are:

Q1: How do the relational methods compare against the propositional methods?

[1] We consider a uniform probability distribution as the initial model, but an expert-designed tree can also be used

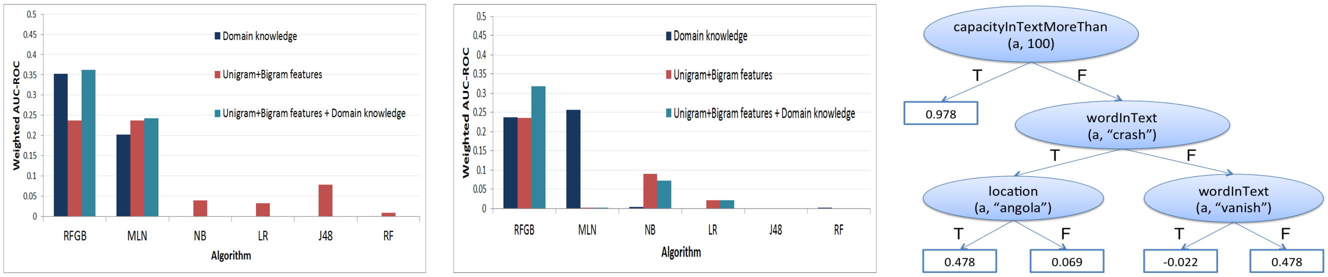

Figure 1: Results of the experiments in: (a) Flight Domain and (b) Literature Domain. (c) A learned tree in the flight domain

Predicate	Explanation
sentenceInDocument(*sID*, *dID*)	*sID* is a sentence in document *dID*
fleschKincaidSentence(*sID*, *val*)	*val* is the *Flesch Kincaid Readability Score* of sentence
gunningFogDocument(*dID*, *val*)	*val* is the *Gunning Fog Readability Score* of document
tfIdf(*dID*, *wID*, *s*)	*s* is the *tf-idf score* of word *wID*
word(*wID*, *wText*)	*wID* is the ID for word *wText*

Table 2: Features for Literature Domain.

Clause

aircraftCapacity(doc, "high") $\longrightarrow$ anomaly(doc)
aircraftLocation(doc, "normal") $\longrightarrow$ anomaly(doc)

gunningFogScore(doc, +dgfs) $\wedge$ fleschScore(doc, +dfs) $\longrightarrow$ AnomalousDocument(doc)
fleschScore(doc, +dgfs) $\wedge$ fleschKincaidScore(doc, +dfs) $\longrightarrow$ AnomalousDocument(doc)

Table 3: Sample MLN rules (top) Flight domain (bottom) Literature domain.

Q2: How useful are the domain knowledge features when compared to the standard unigram and bigram features?

Algorithms Considered: We employ two SRL algorithms and four standard propositional methods for detecting anomalies in the two defined domains. The two relational methods we considered were RFGB and a hand designed MLN. We considered standard machine learning baselines: (1) Logistic Regression, (2) Decision Trees, (3) Naive Bayes and (4) Random Forests. For the SRL algorithms, we considered using: (1) only unigram and bigram features, (2) only domain specific features and (3) all the features. Some sample hand-designed MLN rules are presented in Table 3. For the flight domain (top), we constructed clauses such as *"If the mentioned missing flight is a passenger aircraft with more than 100 passengers, it is an anomaly"*, and *"If the mentioned location is not a war zone it is an anomaly"*. For the literature domain, following previous work, the rules were constructed by simply comparing the readability scores. For example, the first rule in the second half of the table looks for the gunningFogScore and the fleschScore of the document. The second rule uses the fleschScore and the fleschKincaidScore of the document. "+" is a shorthand in MLN softwares for using all the values of the variable. For instance, the first rule, when run through a MLN software such as Alchemy [2] will learn a weight for all combinations of the gunningFogScore (denoted as dgfs) and fleshScore (denoted as dfs). An easy way to understand this is to consider each rule as a template to construct individual instantiated features. These rules, can hence, be interpreted as capturing domain-specific advice which avoid the need to explicitly generate the features. The MLN learning algorithm will then learn parameters (weights) for each of these features. Then the probability of the document is simply a log-linear function of these weighted features. For more details on these scores, we refer to Guthrie (2008).

Evaluation Measures: Standard evaluation techniques on relational models include the use of Area under ROC or PR curves, F1 score etc. rather than just accuracy. However in many domains such as anomaly detection, the model should identify as many positive cases (anomalies) as possible as long as the precision stays within a reasonable range. It is essential that the evaluation metrics assign *higher weights to high recall regions*, that is, the top region in an ROC curve. Following the suggestion of Weng and Poon (2008), we use a weighted-AUC measure by which we divided the ROC curve into five equal parts (N=5), and transferred 80% weight from each region to the one above it and called this as *weighted AUC-ROC*. As with the regular AUC-ROC, higher weighted AUC-ROC indicates better performance.

Results: We ran three-fold cross validation and present the results in Figure 1. As can be seen, in both the domains, the relational methods (RFGB and MLN) in all the three settings outperform the propositional methods significantly when measuring the weighted AUC-ROC. This indicates that relational methods were more effective in identifying the anomalous text. This result is in line with several previously published works of employing these relational learning methods. Hence, Q1 can be answered affirmatively. When comparing the different features, it is clear that in both the domains, the use of domain-specific features helps RFGB's performance. In the flight domain, the number of domain-specific features varied according to the document. It is easier to define a predicate in FOL and allow for mul-

[2] alchemy.cs.washington.edu

tiple instantiations but the same cannot be done easily with the propositional classifiers. Hence, we do not report results for these features for the standard methods.

In the flight domain, the unigram and bigram features were not significant and do not improve the performance even when combined with the domain-specific features. In the literature domain, however, these features are quite effective because of the fact that the anomaly itself is purely text dependent. Hence, grouping all the features there results in a significant increase in performance. For MLNs in flight domain, the rules were not necessarily robust and hence adding unigram and bigram features improve the performance. For the literature domain, the number of unigram and bigram features are quite high and MLN weight learning did not converge even after 48 hours. Following this, Q2 can be affirmatively answered for RFGB but needs more research and investigation for MLNs.

A sample tree for the flight domain is presented in Figure 1.*c*. The left branch states that if the capacity of the flight mentioned in the text of document a is more than 100, then the weight is 0.978 (indicating a high chance of being an anomaly). Else, if the capacity is less than 100 and if the word crash appears in the document along with *angola* as the location, then the weight is 0.478 else it is low with a weight of 0.069. So the flights with high capacity disappearing are rarer than flights crashing in Angola.

We performed a deeper analysis to understand why relational models help. It is clear that in our current experimental set up, we do not consider cross-document learning and hence the relational models do not benefit from such relations. However, the reason for superior performance is the inherent representation of first-order logic - the existential quantifier. Using this quantifier allows us to subsume the potentially infinite feature vector (by taking into account the presence or absence of specific words or features) of the corresponding propositional configuration into a structured relational format. Similar results have been observed in other applications that employ first-order logic.

In summary, our experiments clearly demonstrate the superior predictive nature of the domain-specific features (flight capacities and types in flight domain and lexical features in the literary domain) compared to standard NLP features. They clearly show that SRL models are capable of learning from and exploiting these features effectively.

Discussion

We presented the need for domain specific features and knowledge when identifying anomalies in text documents. We designed two domains - an event anomaly domain and a document anomaly domain and identified the key features in these two domains. Empirical evaluations showed that these features were powerful predictors of anomalies. There are several interesting research directions for future work. Firstly, since we are using relational learning methods, due to the power of existential quantifier, the next natural step is to incorporate a relational structure for these documents and exploit it further to highlight differences between a document with respect to the remaining corpus. The current methods all treat false positives and false negatives equally. While the performance under weighted AUC-ROC of the SRL models are better, modeling the relative costs of false positives and false negatives more faithfully is another direction. Also, identifying anomalies based on jointly learning about multiple events mentioned in the text requires us to significantly extend these methods, a direction which we will pursue. Using more examples either by incorporating weak supervision (Craven and Kumlien 1999) is another future problem that we will consider. Finally, more rigorous evaluation in multiple domains is required.

Acknowledgements

The authors gratefully acknowledge support of the DARPA DEFT Program under the Air Force Research Laboratory (AFRL) prime contract no. FA8750-13-2-0039. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Baker, L.; Hofmann, T.; Mccallum, A.; and Yang, Y. 1999. A hierarchical probabilistic model for novelty detection in text.

Barria, J., and Thajchayapong, S. 2011. Detection and classification of traffic anomalies using microscopic traffic variables. *Intelligent Transportation Systems, IEEE Transactions on* 12:695–704.

Blockeel, H., and Raedt, L. D. 1998. Top-down induction of first-order logical decision trees. *Artificial Intelligence* 101:285–297.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41(3):15:1–15:58.

Craven, M., and Kumlien, J. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*.

Dietterich, T.; Ashenfelter, A.; and Bulatov, Y. 2004. Training conditional random fields via gradient tree boosting. In *ICML*.

Domingos, P., and Lowd, D. 2009. *Markov Logic: An Interface Layer for AI*. San Rafael, CA: Morgan & Claypool.

Getoor, L., and Taskar, B., eds. 2007. *Introduction to Statistical Relational Learning*. MIT Press.

Guthrie, D. 2008. *Unsupervised detection of Anomalous Text*. Ph.D. Dissertation, University of Sheffield.

Mahapatra, A.; Srivastava, N.; and Srivastava, J. 2012. Contextual anomaly detection in text data. *Algorithms* 5:469–489.

Manevitz, L., and Yousef, M. 2002. One-class svms for document classification. *J. Mach. Learn. Res.* 2:139–154.

Manevitz, L., and Yousef, M. 2007. One-class document classification via neural networks. *Neurocomput.* 70(7-9):1466–1481.

Natarajan, S.; Khot, T.; Kersting, K.; Guttmann, B.; and Shavlik, J. 2012. Gradient-based boosting for statistical relational learning: The relational dependency network case. *Machine Learning* 86(1):25–56.

Weng, C., and Poon, J. 2008. A new evaluation measure for imbalanced datasets. In *Proceedings of the 7th Australasian Data Mining Conference - Volume 87*, AusDM '08, 27–32.

Zhang, J., and Zulkernine, M. 2006. Anomaly based network intrusion detection with unsupervised outlier detection. In *Communications, 2006. ICC '06. IEEE International Conference on*, volume 5, 2388–2393.