Event Mining over Distributed Text Streams

John Calvo Martinez
School of Computer Science and Engineering
University of New South Wales
Sydney NSW 2052, Australia
j.calvomartinez@unsw.edu.au

ABSTRACT

This research presents a new set of techniques to deal with event mining from different text sources, a complex set of NLP tasks which aim to extract events of interest and their components including authors, targets, locations, and event categories. Our focus is on distributed text streams, such as tweets from different news agencies, in order to accurately retrieve events and its components by combining such sources in different ways using text stream mining. Therefore this research project aims to fill the gap between batch event mining, text stream mining and distributed data mining which have been used separately to address related learning tasks. We propose a multi-task and multi-stream mining approach to combine information from multiple text streams to accurately extract and categorise events under the assumptions of stream mining. Our approach also combines ontology matching to boost accuracy under imbalanced distributions. In addition, we plan to address two relatively unexplored event mining tasks: event coreference and event synthesis. Preliminary results show the appropriateness of our proposal, which is giving an increase of around 20% on macro prequential metrics for the event classification task.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; *Boosting*; *Distributed algorithms*; • **Information systems → Data stream mining**;

KEYWORDS

Event mining; stream mining; text mining

ACM Reference Format:
John Calvo Martinez. 2018. Event Mining over Distributed Text Streams. In *Proceedings of 11th ACM International Conference on Web Search & Data Mining, Marina Del Rey, CA, USA, February 5–9, 2018 (WSDM 2018),* 2 pages. https://doi.org/10.1145/3159652.3170462

1 INTRODUCTION

The complex task of detecting and extracting events from unstructured text data has been heavily researched over the last decade. Event detection and extraction is a set of NLP tasks which aims to create methods and techniques to identify events, including descriptive elements such as actors, targets, locations, dates, and event types [1]. This information is extracted from unstructured texts such as news articles or social media updates as in Twitter. Researchers have devised several methods and techniques in order to extract such components transforming unstructured information into a more manageable, structured representation to be used by a human analyst e.g. social scientists could benefit by using factual, fine grained information to analyse human rights abuses in violent conflicts [6]. Nonetheless, current event extraction methods are being used on general news text corpora. This is the case for the ACE dataset, which was built to extract detailed events from general news text corpora [1]. However, a fine grained set of event categories naturally increases the imbalanced nature of the data, making it difficult to accurately classify with current state-of-the-art techniques. Pavlick *et al.* [5] recently benchmarked event extraction techniques for the gun violence dataset, a domain specific dataset for gun violence in the US, and low precision and recall results were attained using the best reported techniques in the ACE competition as in [4, 7], reaching 30.2% precision and 20.1% recall on target identification. Furthermore, given that the input can potentially change its underlying statistical distribution at any time and a text stream such as Twitter is frequently updating information near to real time, it is desirable that the learning task is able to predict at any time without losing significant accuracy.

1.1 Research questions

The following research questions are derived from the avobe mentioned challenges:

- How to efficiently categorise text streams under real time conditions and imbalanced distributions?
- How to efficiently recognise event components (actors, targets, locations) from text streams under stream mining constraints?
- How can stream mining techniques be used to find event co-references over multiple sources of information?
- Which stream mining techniques are useful for synthesising events from multiple sources of information?

2 PROPOSED APPROACH

In this section, the techniques to be applied in this research will be described. We propose several stream mining algorithms and techniques applied according to each sub-learning problem.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3170462

2.1 Event type classification

We propose to use a distributed ensemble stream mining approach. This learning algorithm uses the power of domain specific ontologies in order to boost accuracy compared to commonly used methods. During the pre-processing step, feature augmentation is done using the ontology by weighting the relevant words found in the text that matches any category found in the ontology. Afterwards, a stream mining algorithm classifies the incoming instance by combining two techniques, a global Naïve Bayes method and a local matching method. The local method looks at the highest weighted category calculated during the augmentation, and the global learner combines local outputs to classify the incoming instance. We found that a combination between local and global classifiers boosts prequential precision and recall under imbalanced distributions.

2.2 Event component classification

This particular task is being addressed by using a syntactic and semantic feature representation particularly useful for word classification. Contextual features based on a word's syntactic role are proposed, for instance, whether or not the observed word is on the left of the main verb of the sentence, similar to [4] but with more analysis on component selection. The use of such features into an updateable learner seems to improve results for named entities classification (actor, target, location).

2.3 Event coreference resolution

The goal of this particular learning task is to find the closest matches between event references, in order to retrieve a group of text instances reporting the same real event. To achieve this, we propose a stream clustering mechanism using a particular similarity metric we call the "eventiveness" similarity metric, combined with the cosine measure as defined in [8]. Stop words, lemmatization and Porter stemming will be applied before computing the cosine function. Finally, a stream mining model is applied to perform the binary classification task.

2.4 Event synthesis

We propose a multi-stream mining method to combine knowledge from different sources of information. For this purpose, we take all event mentions from each reported event found, including all sources. If a reported component is different across sources, then we use stream mining, using a feature representation of event components, including actors, targets and locations and the information quality measure gathered form each source. In this case, the source of information and its information quality depends only on past true label values. A majority vote algorithm is initially proposed, weighted by the credibility value of each source of information.

3 INITIAL RESULTS

A set of initial results for the event classification task is given in Table 1. Methods were tested on a dataset on the Afghanistan conflict on Twitter during 2016. This ontology was initially used in [3]. In order to compare our results, we use the prequential test-then-train methodology as defined in [2]. Prequential evaluation is complemented by using macro precision and macro recall metrics.

Table 1: Event type classification results

Method		Prequential			Macro Preq.	
		Prec.	Rec.	F1	Prec.	Rec.
TF-IDF	MC	27.8%	85.8%	42%	4.2%	3.8%
	MNB	29.7%	89.7%	55.1%	31.4%	10.1%
Ontology	MNB	42%	**90.5%**	**57.4%**	22.8%	8.8%
	HT	27.3%	88.6%	41.8%	18%	4.9%
	DE(HT)	42.4%	70.1%	53%	51.9%	25.1%
	DE(MNB)	**45.3%**	74.4%	56.5%	**52.1%**	**25.2%**

We used an initial set of existing stream mining algorithms as baselines, including majority class (MC), Multinomial Naïve Bayes (MNB), Hoeffding trees (HT), using both a TF-IDF representation of text and our ontology feature representation. Initial results show an increase in macro precision and macro recall of more than 20% using our distributed ensemble (DE). In addition, we found that very little literature has discussed macro metrics on classification tasks with skewed class distributions.

4 CONCLUSIONS

In this paper, we proposed different methods to address event mining problems under streaming conditions. We propose different techniques for each major event mining task. We focused on improving baselines by using ontology enrichment with a proposed ensemble algorithm. Preliminary results show improvements in macro precision and recall, and more specifically, we have argued that there is a research gap when dealing with imbalanced datasets. Future work will be devoted to improve results in such scenarios and also measuring the efficiency of distributed versions of our algorithms.

ACKNOWLEDGMENTS

The author would like to thank Wayne Wobcke, Alfred Krzywicki, Michael Bain and Bradford Heap for providing support and mentoring to this research and Sandeepa Kannangara for proof-reading. This work is supported by Data to Decisions CRC.

REFERENCES

[1] David Ahn. 2006. The Stages of Event Extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. 1–8.

[2] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. 2010. MOA: Massive Online Analysis. *Journal of Machine Learning Research* 11, 1601–1604.

[3] Bradford Heap, Alfred Krzywicki, Susanne Schmeidl, Wayne Wobcke, and Michael Bain. 2017. A Joint Human/Machine Process for Coding Events and Conflict Drivers. In *Advanced Data Mining and Applications*, G. Cong, W.-C. Peng, W.E. Zhang, C. Li, and A Sun (Eds.). Springer, Cham, 639–654.

[4] Qi Li, Heng Ji, and Liang Huang. 2013. Joint Event Extraction via Structured Prediction with Global Features.. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 73–82.

[5] Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The Gun Violence Database: A New Task and Data Set for NLP.. In *EMNLP*. 1018–1024.

[6] Philip A Schrodt. 2012. *CAMEO: Conflict and Mediation Event Observations Event and Actor Codebook.* Technical Report. Department of Political Science. Pennsylvania State University.

[7] Lei Sha, Jing Liu, Chin-Yew Lin, Sujian Li, Baobao Chang, and Zhifang Sui. 2016. RBPB: Regularization-Based Pattern Balancing Method for Event Extraction.. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1224–1234.

[8] Michael Steinbach, George Karypis, Vipin Kumar, et al. 2000. A Comparison of Document Clustering Techniques. In *KDD Workshop on Text Mining*. 525–526.