Design and Maintenance of Event Forecasting Systems

Sathappan Muthiah

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Narendran Ramakrishnan, Chair
Chang-Tien Lu
Chandan K. Reddy
Ravi Tandon
David Mares

February 10, 2021
Arlington, Virginia

Keywords: Machine Learning, Event Extraction, Forecasting.

Design and Maintenance of Event Forecasting Systems

Sathappan Muthiah

(ABSTRACT)

With significant growth in modern forms of communication such as social media and micro-blogs we are able to gain a real-time understanding into events happening in many parts of the world. In addition, these modern forms of communication have helped shed light into the increasing instabilities across the world via the design of anticipatory intelligence systems [45, 43, 20] that can forecast population level events like civil unrest, disease occurrences with reasonable accuracy. Event forecasting systems are generally prone to become outdated (model drift) as they fail to keep-up with constantly changing patterns and thus require regular re-training in order to sustain their accuracy and reliability. In this dissertation we try to address some of the issues associated with design and maintenance of event forecasting systems in general. We propose and showcase performance results for a drift adaptation technique in event forecasting systems and also build a hybrid system for event coding which is cognizant of and seeks human intervention in uncertain prediction contexts to maintain a good balance between prediction-fidelity and cost of human effort. Specifically we identify several micro-tasks for event coding and build separate pipelines for each with uncertainty estimation capabilities and thereby be able to seek human feedback whenever required for each micro-task independent of the rest.

Design and Maintenance of Event Forecasting Systems

Sathappan Muthiah

(GENERAL AUDIENCE ABSTRACT)

Event forecasting systems help reduce violence, loss/damage to humans and property. They find applicability in supply chain management, prioritizing citizen grievances, designing measures to control violence and minimize disruptions and also in applications like health/tourism by providing timely travel alerts. Several issues exist with the design and maintenance of such event forecasting systems in general. Predictions from such systems may drift away from ground reality over time if not adapted to various shifts (or changes) in event occurrence patterns in real-time. A continuous source of ground-truth events is of paramount necessity for the continuous maintenance of forecasting systems. However ground-truth events used for training may not be reliable but often information about their uncertainty is not reflected in the systems that are used to build the ground truth. This dissertation focuses on addressing such issues pertaining to design and maintenance of event forecasting systems. We propose a framework for online drift-adaptation and also build machine learning methods capable of modeling and capturing uncertainty in event detection systems. Finally we propose and built a hybrid event coding system that can capture the best of both automated and manual event coders. We breakdown the overall event coding pipeline into several micro-tasks and propose individual methods for each micro-task. Each method is built with the capability to know what it doesn't know and thus is capable of balancing quality vs throughput based on available human resources.

Dedications

To my wonderful Mom, Dad, Brothers and Wife

Acknowledgements

First and foremost I would like to thank my advisor Dr. Naren Ramakrishnan for his continued support and encouragement throughout my stay at Virginia Tech. He was not only instrumental in kindling my interest in machine learning and data mining but is also my mentor in many ways. This dissertation would be much less than what it is if not for the thoughftul and insightful comments from my committee. Thank you, Dr. David Mares, Dr. Ravi Tandon, Dr. Chang-Tien Lu and Dr. Chandan Reddy.

I would also like to thank everyone involved with EMBERS project and to all my labmates at the Sanghani Center for Artificial Intelligence and Data Analytics. Mentors like Dr. Prithwish Chakraborty, Dr. Patrick Butler and Dr. Shahriar Hossain at the Sanghani Center have helped shape the early years of my Phd. I have also learnt quite a lot in terms of research skills, perseverance, patience and people skills from Arvinth Chanthar Rathinam, Nathan Self, Nikhil Muralidhar, Dr. Yue Ning and Dr. Rupinder Khandpur. I would also like to extend my sincere thanks to Juanita victoria, Joyce Newberry, Jessica Mullins, Afroze Mohammad and Roxanne Paul for their help with all administrative tasks and for their continued encouragement.

Finally I would like to thank my friends and fellow graduate students Debanjan Datta, Dr. Saurav Ghosh, Sneha Mehta, Subhodip Biswas, Rongrong tao, Taha Hassan, Nurender Chakraborty and Sukrit Venkatagiri for making my stay at Virginia Tech exciting.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern forms of communication such as social media and micro-blogs have helped shed insight into increasing instabilities across the world, e.g., in regions like Latin America and the Middle East. We are now able to better understand and glean deeper insights about the causes of such events. This deluge of information has also led to the development of many anticipatory intelligence systems [45, 43, 20] that can forecast population level events like civil unrest, disease occurrences with reasonable accuracy. Anticipatory intelligence systems or event forecasting systems in general are prone to become outdated (model shift) as they fail to keep-up with constantly changing patterns and thus require regular re-training in order to sustain their accuracy and reliability. Thus, a real-time source of ground-truth information about actual events that occurred is essential.

In this dissertation, we address some of the aforementioned issues. First, we study model drift in event forecasting systems and propose techniques for drift adaptation. Second, we propose a multi-instance learning based approach capable of modeling and capturing uncertainty with applications to event detection. Finally, we build a human-aided automated system for event encoding which is cognizant of and seeks human intervention in uncertain prediction

contexts to maintain a good balance between prediction-fidelity and cost of human effort. In the following sections, we setup the problem and discuss in detail the sub-problems studied.

1.1 What is an event of interest?

We define an event of interest to be any significant societal event that has happened in the recent past and is reported in a national news paper of repute. Societal events of interest include *Military Action*, *Non-state Actor* and *Civil Unrest* events. An event record generally contains information regarding — *who, when, where and why.* For example, a Military Action event is represented by a tuple $<$*Actor, Target, type, Location, Date* $>$. An example of a Military Action event is shown in Figure. 1.1.

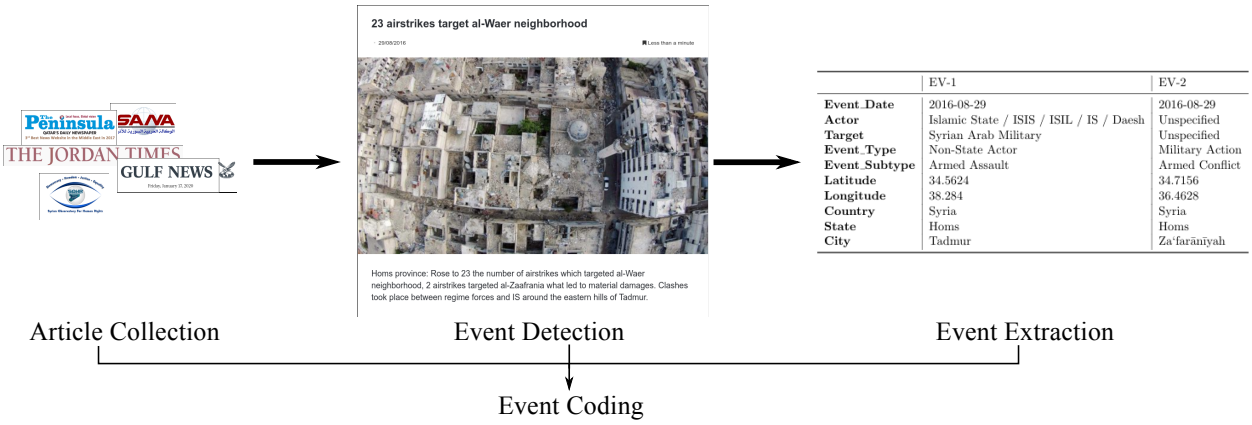
1.2 What is event coding?

Event coding refers to the process of obtaining a structured representation of an ongoing (or recently concluded) event reported in newspapers. This structured representation includes information about the location (at city level granularity), Actor (or Perpetrator), Target (or Victim), Reason, Type (civil unrest, military action etc.,) and date of occurrence of the event. Figure 1.1 provides an example of event coding. The figure showcases the different steps involved in event coding like – 1) Article collection, wherein news articles from reputed national news papers are collected on a daily basis, 2) Event detection which involves identifying if an article is reporting an event of interest and finally 3) Event extraction referring to the process of extracting all necessary meta-information about an event from the article text.

1.3 Motivation and Organization of the Dissertation

Event forecasting in general has several use cases and can help reduce violence, loss/damage to humans and property. Event forecasting finds applicability in supply chain management, prioritizing citizen grievances, designing measures to control violence and minimize disruptions and also in applications like health/tourism by providing timely travel alerts. However such systems require regular maintenance and are prone to model drift. Also a real-time source of ground-truth information is of paramount importance in several application areas like disease surveillance, intelligent systems and governance so as to maintain quality and reliability. Most existing event coders are either not scalable or are un-reliable and unable to keep up with the rapid increase in terms of data availability. Existing automated event coders [43, 30] still make use of traditional dictionary-based approaches and do not leverage recent advances in machine learning.

In this dissertation our goal is to address some of the issues regarding maintenance of an event forecasting system namely — 1) model drift, 2) uncertainty and finally, 3) continuous source of ground-truth event data to keeping the system operational. First we describe our experiences in operating a real-time 24x7 anticipatory intelligence system to understand the importance of timely and reliable ground-truth generation along with the implications of model drift associated with running the system for a long period of time. With this observation we then propose techniques for performing uncertainty-aware event detection and finally we build a hybrid event coding framework composed multiple machine learning pipelines to extract the different subsets (or features) of the ground-truth.



	EV-1	EV-2
Event_Date	2016-08-29	2016-08-29
Actor	Islamic State / ISIS / ISIL / IS / Daesh	Unspecified
Target	Syrian Arab Military	Unspecified
Event_Type	Non-State Actor	Military Action
Event_Subtype	Armed Assault	Armed Conflict
Latitude	34.5624	34.7156
Longitude	38.284	36.4628
Country	Syria	Syria
State	Homs	Homs
City	Tadmur	Zaʿfarānīyah

Figure 1.1: An example of the event coding where in from a news paper article is identified to report an event of interest and 2 events, one Military Action event and one Non-State Actor event, are extracted

1.4 Problem 1: Studying Model-drift in Event Forecasting Systems

Forecasting societal events such as civil unrest has a long tradition in the intelligence analysis and political science community. In this problem, we study the deployment of EMBERS (Early Model Based Event Recognition using Surrogates) system [45], an anticipatory intelligence system [14] that forecasts significant societal events (e.g., Civil unrest, Military actions, and Non-state Actor based events.). We look at the discoveries it has enabled, and lessons learned including our perspectives on the limits of forecasting and ethical considerations. In particular, we provide detailed insights about the value proposition to an analyst and how EMBERS forecasts are communicated to its end-users. We identify that to successfully operate such a forecasting system 24x7, we require a constant source of ground-truth to keep the models up-to-date and prevent model drift. During the operation of EMBERS, a constant source of Gold Standard Reports (GSR), a monthly catalog of events as reported in newspapers of record, was compiled by human analysts at the MITRE corporation.

1.5 Problem 2: Modeling and Capturing Uncertainty in Event Detection

In this problem, we tackle the first step of event coding, i.e., event detection. Based on our experiences we identify that a continuous source of event data is paramount for maintaining the reliability and success of an event forecasting system. Event detection is the process wherein a decision is made regarding whether a news article reports an ongoing or just completed event of interest (like Civil unrest, Military Action). This step is very important in event coding as it helps reduce significantly the articles that pass onto underlying and more computationally intensive steps. Specifically we focus on modeling the uncertainty at a fine-grained level so that the model knows what it doesn't know.

1.6 Problem 3: Hybrid Micro Tasking Framework for Event Coding

In the previous problem we introduced a system for identifying which articles report an event of interest. Once such articles are identified we need a framework for extracting necessary meta-information about an event(or events) from the article. We try to build such a framework in this part of the dissertation. Most traditional event coding frameworks are either entirely automatic [43] or are fully coded by human analysts. Wei et.al., [30] showed that fully automatic systems lack significantly as compared to human annotated systems. We thus build a hybrid event extraction system that is capable of seeking human intervention for certain sub-tasks (called micro-tasks) when it is uncertain of its prediction (or extraction).

1.7 Related Work

1.7.1 Event Coding

Hogenboom et al. [25] provides an overview of different extraction methodologies (statistical and linguistic) used by the current state-of-the-art systems. Schrodt et al. [48] introduced one of the earliest event coders TABARI (Textual Analysis by Augmented Replacement Instructions). It searches for hand coded patterns on only the first few sentences of a news article to encode an event of interest. TABARI was succeeded by JABARI and PETRARCH [49]. BBN's SERIF (statistical Entity and Relation Information Finder) is another state-of-the-art event coder using several NLP components to identify triplets of type actor-subject-target from article text (at both sentence and document level). It is to be noted that the SERIF encoder is not available for public use. The ICEWS (Integrated Crisis Early Warning System) [5] makes use of the SERIF encoder to parse hundreds of articles to produce a real-time database of events. GDELT(Global Database of Events, Language and Tone) [33] another event coding system that churns out events from a larger geographical area and categories makes use of an enhanced version of the TABARI parser. The above mentioned systems are fully automatic and were found to be unreliable as shown be Wei et al. [30]. In order to bridge the gap between fully autonomous systems and highly reliable but extremely costly manual event coders systems. Parang et al. [47] proposed a semi-automated system called EMBERS AutoGSR. In the the EMBERS AutoGSR, several recommendation models were used to help reduce the overall time spent per article by human analysts .

1.7.2 Uncertainty in classification (I don‘t know classification)

In order to support real-word decision making we want the classifier to be able to say "i don't know" or ascertain low confidence when presented with difficult and out-of-distribution instances. Kendall et al. [29] enlist two two major kinds of uncertainties we need to model – 1) Aleatoric uncertainty and 2) Epistemic uncertainty. Aleatoric or data uncertainty refers to uncertainty stemming from the inherent data noise, the known unknowns and is generally unreducible. On the other hand Epistemic or Model uncertainty refers to the uncertainty in model parameters conditional on training data. Generally model uncertainty can be reduced with increase in training data size. Malinin et al. [38] talks about a third kind of uncertainty called distributional uncertainty, the unknown unknowns, which refers to the uncertainty stemming from out-of-distribution data. Generally data uncertainty has been handled via three main strategies - 1) learning classifiers with an inbuilt reject option [2, 53, 35, 18], 2) improving confidence calibration [44, 41, 17, 22] and 3) using the dempster-shafer theory of evidence [21] as shown in sensoy et al. [50]. Model uncertainty can be handled via ensemble modeling as in [16, 56, 58].

Chapter 2

Studying Model-drift in Event Forecasting Systems

In this chapter we study how we can identify model drift in event forecasting systems and propose techniques for drift adaptation. Specifically we detail our experience in running a real-time 24x7 event forecasting system. We provide examples of both successes and misses of the system and provide a structured view of the uncertainties involved in forecasting. Ethical issues associated with such systems is also presented.

2.1 Introduction

In KDD 2014, EMBERS [45], a deployed anticipatory intelligence system [14] that forecasts significant societal events (e.g., civil unrest events such as protests, strikes, and 'occupy' events) using a large set of open source indicators such as news, blogs, tweets, food prices, currency rates, and other public data was introduced. The EMBERS system has been running continuously 24x7 for nearly 4 years at this point and our goal here is to present the

discoveries it has enabled, both correct as well as missed forecasts, and lessons learned from participating in a forecasting tournament including our perspectives on the limits of forecasting and ethical considerations. In particular, we shed insight into the value proposition to an analyst and how EMBERS forecasts are communicated to its end-users.

The development of EMBERS is supported by the Intelligence Advanced Research Projects Activity (IARPA) Open Source Indicators (OSI) program. EMBERS currently focuses on multiple regions of the world but for the purpose of this paper we focus primarily on the 10 Latin American countries, specifically the countries of Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela. Similarly, EMBERS generates forecasts for multiple event classes—influenza-like-illnesses [8], rare diseases [46], elections [37], domestic political crises [30], and civil unrest—but in this paper we focus primarily on civil unrest as this was the most challenging event class with hundreds of events every month across the countries studied here. EMBERS forecasts are scored against the Gold Standard Report (GSR), a monthly catalog of events as reported in newspapers of record in these 10 countries. The GSR is compiled by MITRE corporation using human analysts.

Our key contributions can be summarized as follows:

1. Unlike retrospective studies of predictability, EMBERS forecasts are communicated in real-time before the event to MITRE/IARPA and scored independently of the authors. (Further, the scoring criteria are set by IARPA.) We present multiple quantitative indicators of EMBERS performance as well as insights into how we made EMBERS forecasts most valuable to analysts. We report two primary ways in which analysts utilize EMBERS and present the use of automated narratives to help make EMBERS forecasts as useful as possible.

2. In an attempt to demystify the state-of-the-art in forecasting and to create an open dialogue in the community, we report both successful forecasts of EMBERS as well as events missed by EMBERS. The events not forecast by EMBERS lead us to considerations of both the limitations of the underlying technology as well as the inherent limits to forecasting large-scale events.
3. While social media is often touted as the key to event forecasting systems such as EMBERS, we present the results of an ablation study to outline the performance degradation that ensues if data sources such as Twitter and Facebook were to be removed from the forecasting pipeline.
4. We consider the separation of civil unrest events into events that happen with a degree of regularity versus rare or significant events, and evaluate the performance of EMBERS in forecasting such surprising events.
5. We describe our current best understanding of the limitations to forecasting civil unrest events using technologies like EMBERS and also consider the ethical considerations of the EMBERS technologies.

2.2 Background

We begin by providing a brief review of forecasting systems, followed by a quick preview of EMBERS, its system architecture, machine learning models, and measures for evaluating its performance. For more details, please see [45].

Forecasting societal events such as civil unrest has a long tradition in the intelligence analysis and political science community. We distinguish between forecasting systems versus event coding systems (systems that provide structured representations of ongoing events reported

in newspapers), and focus on the former. Early forecasting systems such as ICEWS [43] provided very broad coverage in countries but were limited by their spatio-temporal resolution (e.g., typically country- and month- level forecasting for specific events of interest [55]). The ICEWS events of interest are domestic political crises, international crises, ethnic/religious violence, insurgencies, and rebellion. A similar project in scope is PITF (Political Instability Task Force) [19] funded by the CIA. To the best of our knowledge, only EMBERS provides the most-specific spatial resolution (city-level) and the most-specific temporal resolution (daily-level) capability in forecasting.

The software architecture of EMBERS (Early Model Based Event Recognition using Surrogates) is designed as a loosely coupled, share-nothing, highly distributed pipeline of processes connected via ZeroMQ. In this manner, the system is both highly scalable and fault tolerant. The EMBERS pipeline can loosely be broken up into four stages: ingestion, enrichment, modeling, and selection. In the first stage, ingestion, data is collected from a variety of sources and streamed into the following stages in real-time. The enrichment stage takes the raw data from the ingestion stage and processes it in various ways including natural language processing, geocoding, and relative time phrase normalization. After enrichment, the modeling stage feeds the enriched data into the various models that make up EMBERS. Unlike other systems which use single monolithic models to make predictions, EMBERS combines the results of several different models to arrive at the most accurate forecasts. In particular, the separate alerts from each model are de-duplicated, fused, and selected and finally emitted as a full forecast for a real world event.

The structure of a civil unrest forecast is shown in Figure 2.1 (left). A forecast constitutes four fields, corresponding to the when, where, who, and why of the protest. These fields are respectively denoted as the date, location, population, and event type. Location is recorded at the city level. Population and event type are fields chosen from a categorical set of

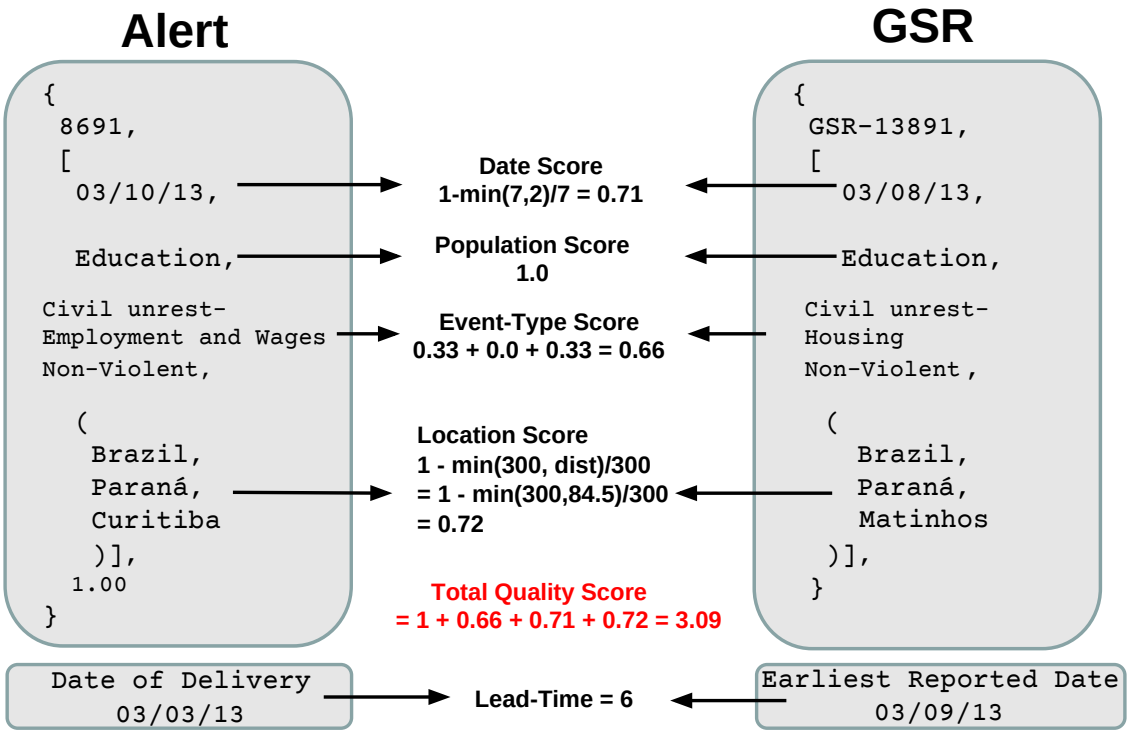


Figure 2.1: An example depicting how an alert is scored with respect to the ground truth.

possibilities. The figure further shows how an alert with all these fields are scored against a GSR event. In the basic scoring methodology shown in Figure 2.1 each of the four fields are weighted uniformly and a total quality score out of 4 is obtained. Apart from this each alert also has a lead-time associated with it calculated as shown in Figure 2.2.

Rather than design one model to integrate all possible data sources, EMBERS adopted a multi-model approach to forecasting. Each model utilized a specific (possibly overlapping) set of data sources and is tuned for high precision, so that the union of these models can be tuned for high recall. A fusion/suppression engine [24] allows a tunable strategy to issue more or fewer alerts depending on whether the analyst's objective is to obtain a higher precision or recall. The underlying models used in EMBERS are: (i) *planned protest model* [39], (ii) *dynamic query expansion* [60], (iii) *volume-based model* [31], (iv) *cascade regression* [6], and (v) a baseline model. The planned protest model, for news and social media (Twitter,

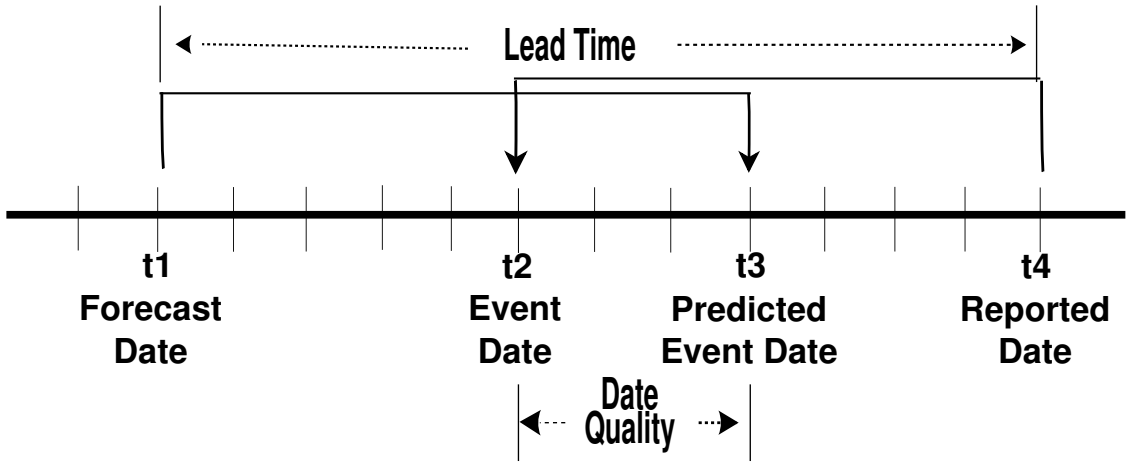


Figure 2.2: Alert sent at time $t1$ predicting an event at time $t3$ can be matched to a GSR event that happened at time $t2$ and reported at time $t4$ if $t1 < t4$.

Facebook), identifies explicit signs of organization and calls for protest, resolves relative mentions of time (e.g., 'next Saturday') and space (e.g., 'the square') to issue forecasts. The dynamic query expansion (DQE) model uses Twitter as a data source and learns time- and country-specific expansions of a seed set of keywords to identify specific situational circumstances for civil unrest. For instance, in Venezuela (an economy where the government exercises stringent price controls), there were a series of protests in 2014 stemming from the shortage of toilet paper, a novel circumstance that was uncovered by DQE. The volume-based model uses a range of data sources, spanning social, economic and political indicators. It uses classical statistical models (LASSO and hybrid regression models) to forecast civil unrest events using features from social media (Twitter and blogs), news sources, political event databases (ICEWS and GDELT [32]), Tor [12] statistics, food prices, and currency exchange rates. It aims to provide a multi-source perspective into forecasting by leveraging the selective superiorities of different data sources. The cascade regression model aims to model activity related to organization and mobilization in Twitter [6]. Finally, the baseline model uses maximum likelihood estimation over the GSR to issue history-based forecasts.

The EMBERS project is unique not just in its algorithmic underpinnings but also in the use of new measures for evaluation, specifically aimed at determining forecasting performance. As shown in Figure 2.2, one of the primary measures of EMBERS performance is lead time, the number of days by which a forecast 'beats the news', i.e., the date of reporting of the event. Lead time should not be confused with date quality, i.e., the difference between the predicted date and the actual date of the event. The date quality is one of the components of the quality score, the other components being the location score, event type score, and population score. Figure 2.1 shows how these other components are scored between an EMBERS forecast and a GSR record.

Given a set of alerts and a set of GSR events for a given month, the lead time is used as a constraint to define legal (alert, event) pairs so that we can construct a bipartite matching to optimize the best quality score. From this bipartite matching, measures of precision and recall can be derived, i.e., by assessing the number of (un)matched events or alerts. Finally, a confidence score is used to assess the quality of probabilities imputed by EMBERS to its forecasts, and measured in terms of the Brier score. For more details, please see [45].

We now turn to a discussion of specific discoveries enabled by EMBERS, into civil unrest in Latin America, and into the complexity of the forecasting enterprise as a whole.

2.3 Performance Analysis

First, we begin with a performance analysis of EMBERS, from both a quantitative point of view with respect to the GSR and with respect to end-user (analyst) goals.

Targets

	Month 12	Month 24	Month 36
Mean Lead-Time	1 day	3 days	7 days
Mean Probability Score	0.60	0.70	0.85
Mean Quality Score	3.0	3.25	3.5
Recall	0.50	0.65	0.80
Precision	0.50	0. 65	0.80

Actual

Metric	Month 12	Month 24	Month 36
Mean Lead-Time	3.89 days	7.54 days	9.76 days
Mean Probability Score	0.72	0.89	0.88
Mean Quality Score	2.57	3.1	3.4
Recall	0.80	0.65	0.79
Precision	0.59	0.94	0.87

Figure 2.3: IARPA OSI targets and results achieved by EMBERS.

2.3.1 Quantitative Metrics

Figure 2.3 depicts both the targets set by the IARPA OSI program as well as the actual measures achieved by the EMBERS system. As shown here, the easiest target to achieve in EMBERS was, surprisingly, the lead time objective. This was feasible due to EMBERS's focus on modeling both planned and spontaneous events. Planned events are sometimes organized with as many as several weeks of lead time and thus identifying indicators of organization was instrumental in achieving lead time objectives. The confidence (mean probability) scores were also achieved by EMBERS and involved careful calibration of probabilities by taking into account estimates of model propensities and data source reliabilities. The measure that was most difficult to achieve was the quality score as it involved a four component additive score and thus tangible improvements in score required more than incre-

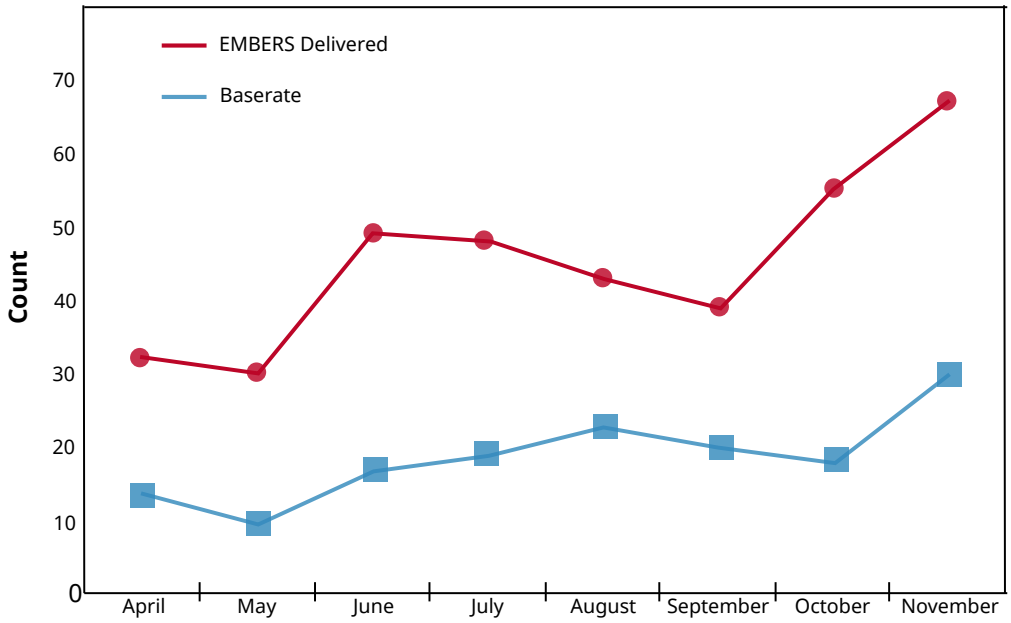


Figure 2.4: Comparison of number of perfect scores (4.0) obtained by EMBERS vs a baserate model each month in 2013.

mental improvements in forecasting specific components. Finally, recall and precision involve a natural underlying trade-off and the deployment of our fusion/suppression engine provided the ability to balance this trade-off to meet IARPA OSI's objectives. Apart from comparing mean scores another interesting measure is to see how many perfect matches (4.0 quality score) are obtained by EMBERS. Fig. 2.4 shows the number of alerts issued by EMBERS that matched perfectly to an event in the future on a monthly basis for 2013. It is clear that EMBERS makes almost double the number of fully accurate forecasts as compared to a baserate model.The baserate model generates alerts using the rate of occurrence of events in the past three months.

2.3.2 Analyst Evaluation

In addition to the quantitative measures above, our experience interacting with analysts demonstrated an interesting dichotomy as to how analysts use EMBERS alerts. Some an-

EMBERS forecasts that there will be a violent protest on February, 18th 2014 in Caracas, the capital city of Venezuela. It predicts that the protest will involve people working in the business sector. The protest will be related to discontent about economic policies.
There were 5, 5, and 5 other similar warnings in last 2, 7 and 30 days, respectively.
The forecast date of the warning falls in week 7, which may have historical importance; this week is found to be statistically significant (pval=0.00461919415894, zscore=2.832, avg. count=57.25, mean=21.569 +/- 12.597)
Audit trail of the warning includes an article printed 2014-02-17.
Major players involved in the protest include Venezuelan opposition leader, students, President Nicolas Maduro, and Leopoldo Lopez.
Reasons: Protest against rising inflation and crime; Protestors want a political change; President Nicolas Maduro has accused US consular officials and right-wing.
Protests are characterized by: Venezuelan opposition leader spearheaded days of protest and calling for peaceful demonstration; Maduro accused official on 2014-12-16; Protests have seen several deadly street protests; Three people were killed on 2014-02-12; Demonstrations setting days of clashes; supporters to march to Interior Ministry on 2014-02-18.

Figure 2.5: An example narrative for an EMBERS alert. Here, color red indicates named entities, green refers to descriptive protest related keywords. Items in blue are historical or real time statistics and those in magenta refer to inferred reasons of the protest.

alysts preferred to use EMBERS in an 'analytic triage' scenario wherein they could tune EMBERS for high recall so that they would apply their traditional measures of filtering and analysis to hone in on forecasts of interest. Other analysts instead viewed EMBERS as a data source and preferred to use it in a high precision mode, e.g., wherein they were focused on a specific region of the world (e.g., Venezuela) and/or aimed to investigate a particular social science hypothesis (e.g., whether disruptions in global oil markets led to civil unrest).

To support these diverse classes of users, we implemented two mechanisms in the alert delivery stage. First, we implemented a mechanism wherein in addition to generating alerts, EMBERS also forecasted the expected quality score for each forecast (using machine learning methods trained on past GSR-alert matches). This expected quality score measure provided a way for analysts to use quality directly as a way to tune the system to receive greater or fewer alerts. Second, we implemented an automated narrative generation capability (see Fig. 2.5) wherein EMBERS auto-generates a summary of the alert in English prose. As shown

in Fig. 2.5, a narrative comprises many parts drawn from different sources of information. One source constitutes the named entities wherein the system uses 'Wikification' to identify definitions and descriptions of named entities on Wikipedia. A second source is historical (or real-time) statistics of warning output and warning performance and situating the alert in this context. The third source pertains to inferred reasons for the protest using knowledge graph identification techniques.

2.4 EMBERS Successes and Misses

Next, we detail some of the successful as well as not so successful forecasts made by EMBERS over the past few years in Latin America.

2.4.1 Successful Forecasts

Brazilian Spring (June 2013)

These protests were the largest and most significant protests in Brazil's recent history and caught worldwide attention. Millions of Brazilians took part in these demonstrations, also known as the Brazilian Spring or the Vinegar Movement (inspired from the use of vinegar soaked cloth by demonstrators to protect themselves from police teargas). These protests were sparked by an increase in public transport fares from $R\$3$ to $R\$3.20$ by the government of President Dilma Rousseff.

As shown in Fig. 2.6, while missing the initial uptick, EMBERS did forecast the increase in the order-of-magnitude of protest events during the Brazilian Spring and also captured the spatial spread in the events. In addition EMBERS correctly forecast that this event will span the broad Brazilian general population (as opposed to being confined to specific

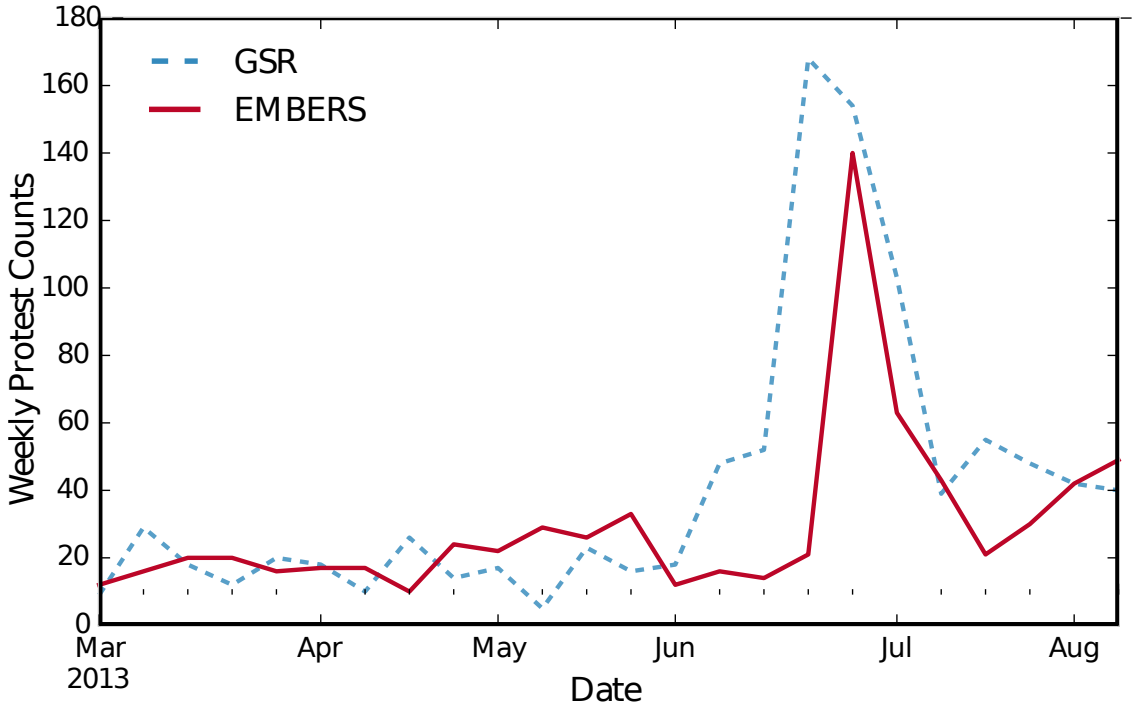


Figure 2.6: EMBERS performance during the Brazilian Spring (June 2013).

sectors).

Around 68% of EMBERS alerts during this period originated from the planned protest model. This is due to the fact that social networking platforms (Twitter and Facebook) as well as conventional news media played a key role in organization of these uprisings. Although initial protests were primarily due to the bus fare increases, they quickly morphed into more broader dissatisfaction to include wider issues such as government corruption, over-spending, and police brutality. The demonstrators also made calls for political reforms. In response, President Rousseff proposed a plebiscite on widespread political reforms in Brazil (but this was later abandoned). Through its dynamic query expansion model, EMBERS was able to capture such discussions on Twitter (see Fig. 2.7), and tracked their evolution as events unfolded through June.

The protests intensified in late June (see Fig. 2.6), which were forecast correctly by EMBERS,

Figure 2.7: Word cloud representing tweets identified by EMBERS dynamic query expansion model.

and these events also coincided with FIFA 2013 Confederation Cup matches. We believe this was an important factor in helping the protests gain momentum, as the events were covered by international media. A majority of protests occurred in the cities that were hosting FIFA soccer matches. EMBERS issued most of its alerts for these host cities (see Fig. 2.8), viz. Rio de Janeiro, São Paulo, Belo Horizonte, Salvador, and Porto Alegre, among others. For example, on 27th June during the Confederations Cup semi-final in Fortaleza, around 5000 protesters clashed with the police near the Castelao stadium. In this case EMBERS had forecast an alert the day before. Later on the 30th of June, when the last games of the confederation cup took place in Rio de Janeiro and Salvador, they were plagued by mass protests as well; EMBERS predicted these events and submitted multiple alerts for Rio for the 28th and 29th June and one for Salvador for 29th June.

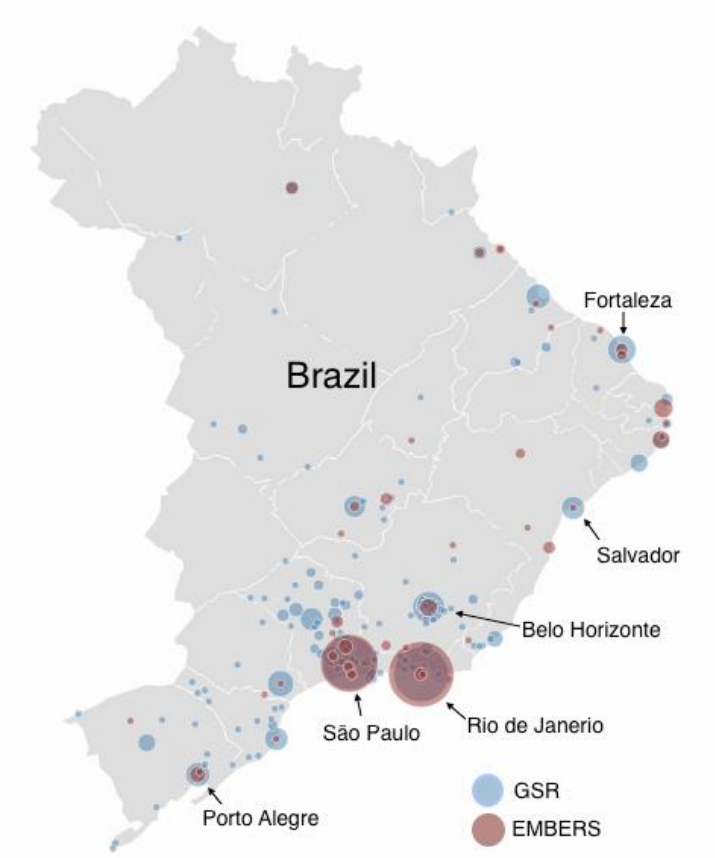


Figure 2.8: Geographic overlap of protest events (from the GSR) and EMBERS alerts for Brazil during June 2013.

Venezuelan protests (Feb-March 2014)

In early 2014 Venezuela began experiencing a situation of turmoil with a large portion of its population protesting due to insecurity, inflation and shortage of basic goods. This period saw one of the highest levels of civil disobedience in Venezuela with protests beginning in January with the murder of a former Miss Venezuela. However, the protests started gaining more importance and turned violent and more frequent with students joining the movement following an attempted rape of a student on a campus in San Cristobal. EMBERS captured some of these first calls-to-protest at San Cristobal and its nearby surrounding areas and correctly forecast the population (Education) and that the protests would turn violent. A majority of the protesters were demanding that president Nicolas Maduro step down owing to the poor economic policies and widespread corruption. EMBERS succeeded in capturing that the reason behind the protests were mainly against government policies with corruption being a major theme. The EMBERS models working on Twitter were also clearly able to identify some of the major leaders involved in the protest, such as the key opposition leader Leopoldo Lopez. Though the events mainly began in San Cristobal, they spread widely throughout the country; EMBERS captured this spread very well as shown in Fig. 2.9. Fig. 2.10 shows how EMBERS closely forecast the spike in the number of events during this period.

Mexico protests (Oct 2014)

In September 2014, there were some peaceful protests by students from Ayotzinapa in Mexico against discriminatory hiring practices for teachers. During these protests, police opened fire on the students killing around three; 43 students went missing. This poor handling of the protest by the Mexican government caused widespread demonstrations throughout the

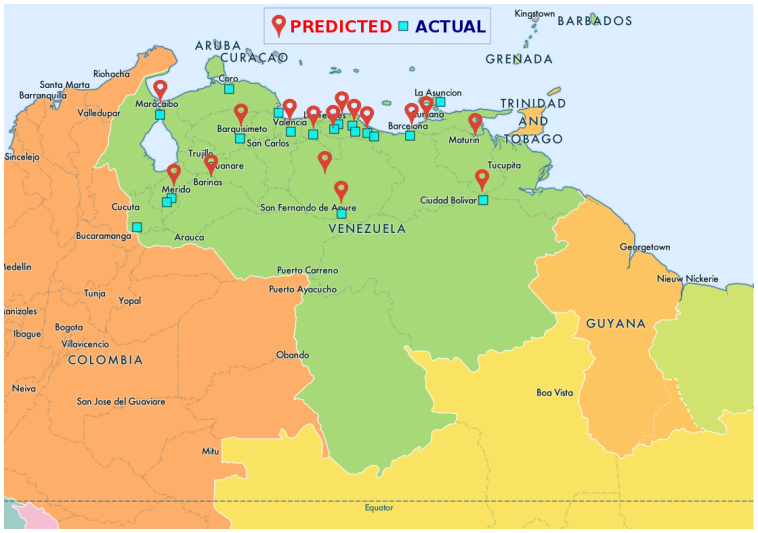


Figure 2.9: Geographical spread of protests (and forecasts) during the Venezuelan student protests (Feb-Mar 2014).

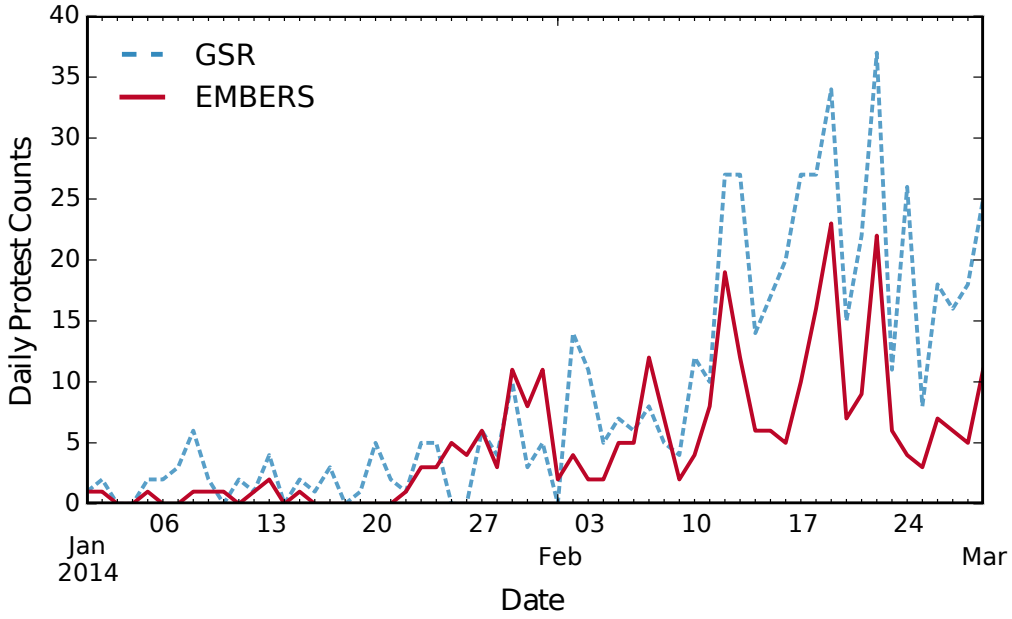


Figure 2.10: EMBERS performance during the Venezuelan student protests (Feb-Mar 2014).

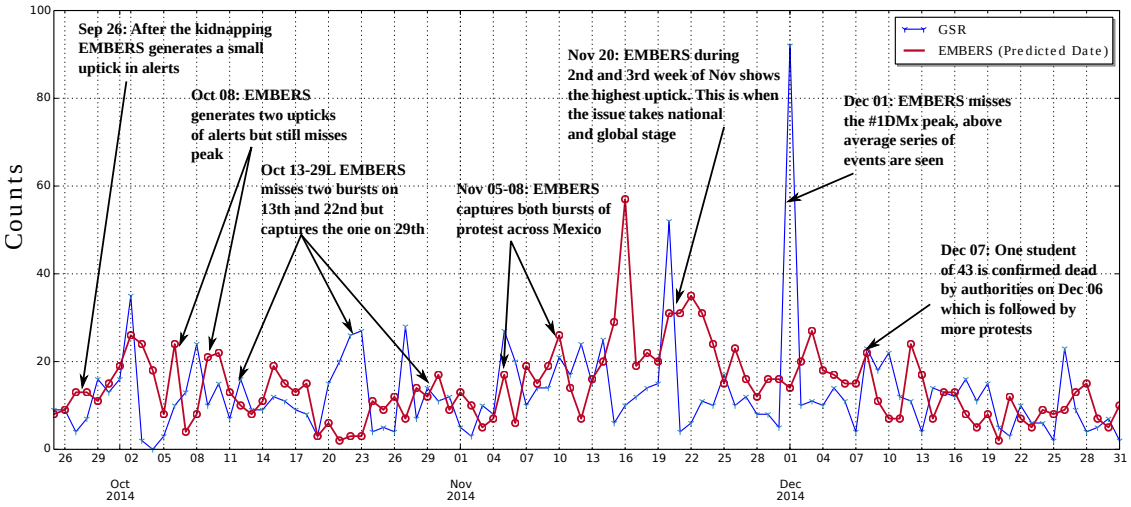


Figure 2.11: Timeline of Mexico protests, showing the correspondence between counts of GSR events and EMBERS alerts on a daily basis.

country over the next few months in support of the families of the 43 missing students. Many of these protests were violent in nature with demonstrators expressing extreme dissatisfaction against the government of president Pena Nieto. EMBERS, as shown in Fig. 2.16 forecast an uptick in Mexico protests during early October 2014 with a lead time of about three days. It also generated a series of alert spikes coinciding with the first large-scale nationwide protests between October 5th and 8th. Fig. 2.11 provides a timeline of GSR events and EMBERS alerts for Mexico during this period. This figure provides a detailed comparison of the continuous stream of alerts produced by EMBERS during this period against how the actual events unfolded in the real world.

Colombia protests (Dec 2014 to March 2015)

Colombia witnessed two different significant protests during this period, one during late December 2014 and the other during February 2015. Towards the end of 2014, the Colombian government was on the process of moving forward with peace negotiations to end 50 years of conflict with the Revolutionary Armed Forces of Colombia (FARC). With the FARC rebels

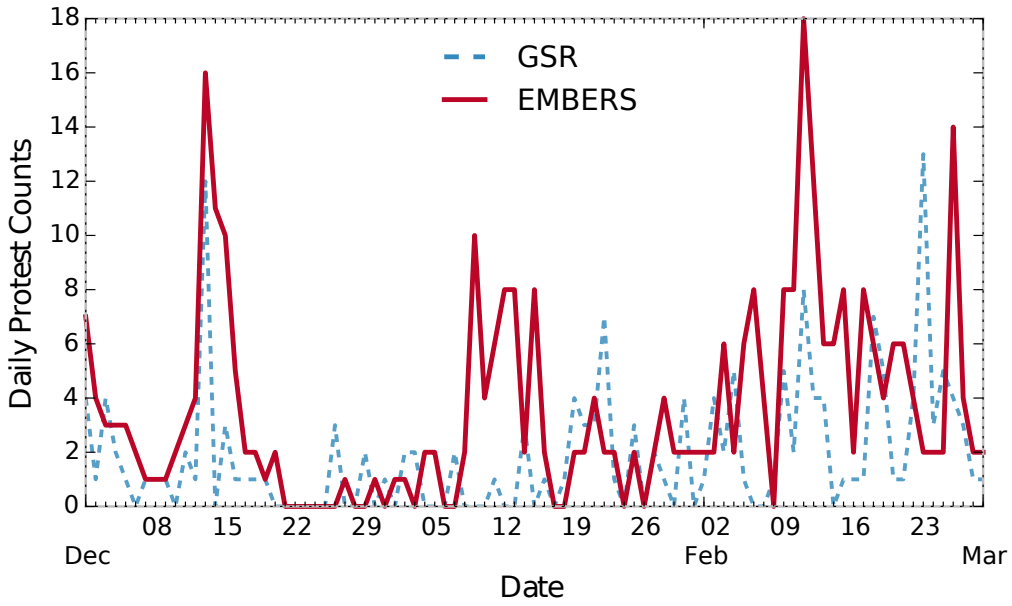


Figure 2.12: EMBERS performance during the Colombia protests (Dec 2014 to Mar 2015).

having been associated with various acts of terror, e.g., extortion, armed conflict, kidnapping, ransom, and illegal mining, for a long period, the people of Colombia gathered in huge numbers to protest against possible amnesty for the FARC rebels. EMBERS successfully forecast the uptick in the number of events during the middle of December 2014 as indicated in Fig. 2.12. This figure also depicts the increase in protest counts during February 2015, although in this case EMBERS over-predicted the counts.

The protests in February 2015 were qualitatively different in nature and were led by truckers unions demanding better freight rates, labor rights, and revolting against high fuel prices. The truckers protests extended for about a month and caused an estimated loss of about $300 million to the Colombian economy. EMBERS forecast the truckers protests accurately at the onset of these events but over-estimated the number of protests during February 11-12.

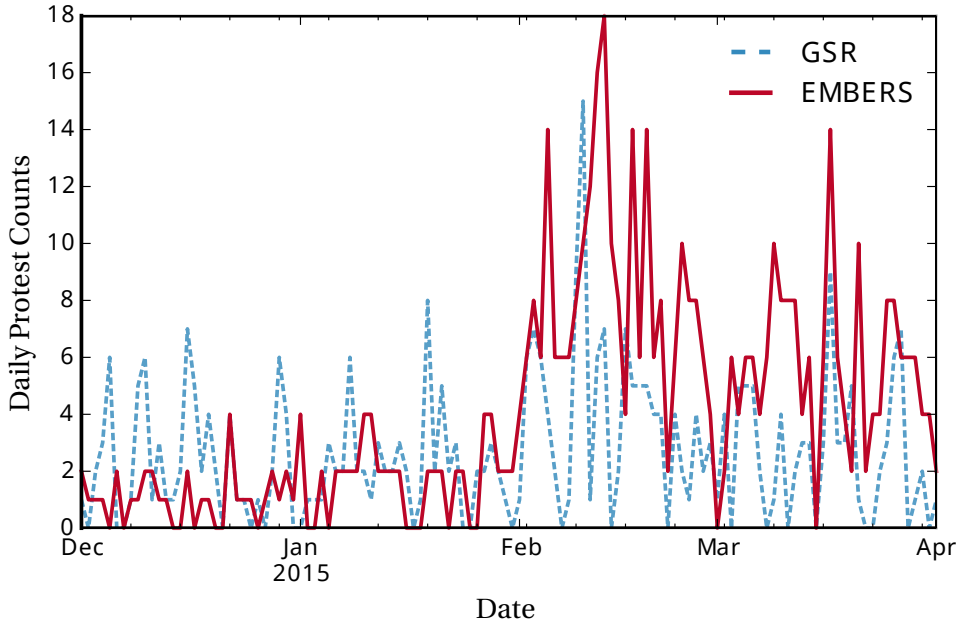


Figure 2.13: EMBERS performance during the Paraguay protests (Feb. 2015).

Paraguay protests (February 2015)

The February 2015 protests in Paraguay were mainly carried out by peasants against the actions of President Horacio Cartes. The protests were carried out after president Horacio's public revelation that he had opened two private Swiss bank accounts. The protests also had a historical significance. They were also being carried out as a tribute to peasant leaders and activists who were murdered. The peasants also protested against the introduction of a new public-private partnership law. EMBERS forecast the uptick in the number of Paraguay protests during mid February 2015 as shown in Fig. 2.13.

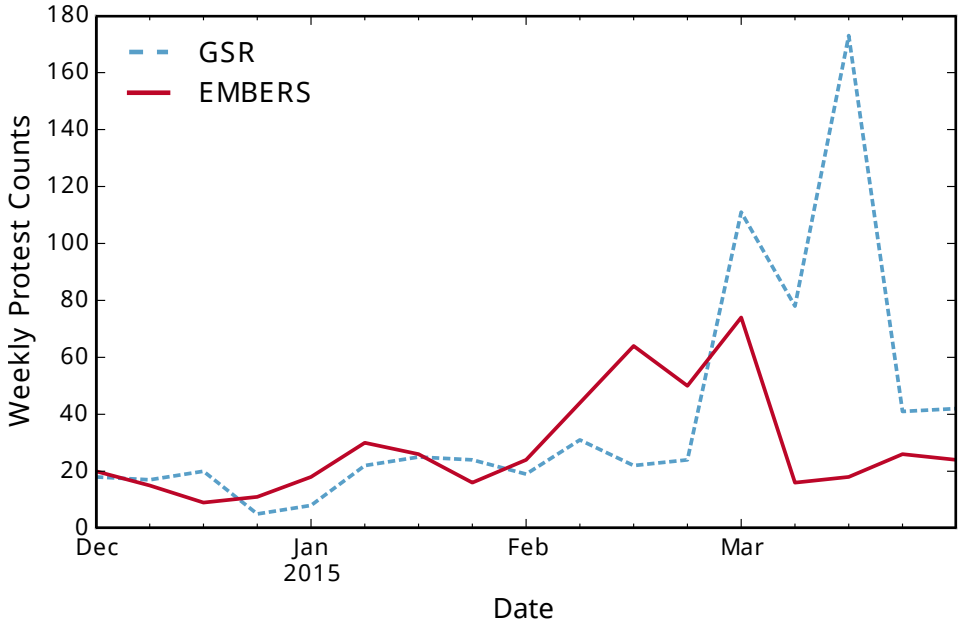


Figure 2.14: EMBERS performance during the Brazilian protests of March 2015.

2.4.2 EMBERS Misses

Next, we outline specific large-scale events that EMBERS failed to forecast accurately, along with a discussion of underlying reasons.

Brazilian protests (March 2015)

The beginning of 2015 saw a series of protests in Brazil demanding the removal of president Dilma Roussef amidst much furore against the increasing corruption in the country. The number of protests increased significantly due to the revelations that many politicians belonging to the ruling party accepted bribes from the state-run energy company Petrobas. The protests drew huge participation from the general population, with protesters generally estimated to be around a million. EMBERS, as shown in Figure 2.14, picked up the onset of events but failed to capture the sudden rise in the number of events.

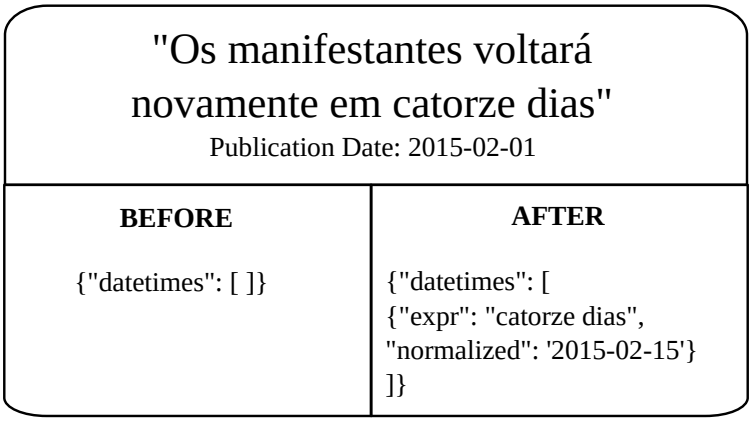


Figure 2.15: Example depicting the improvement in time phrase recognition after changes to Heideltime.

During this period there was a significant architectural change in the EMBERS processing pipeline. As mentioned in Section 2.2 the EMBERS system enrichment pipeline consists of the following steps: natural language processing, geocoding, and relative time phrase normalization (temporal tagging). During early 2015 EMBERS had moved to the Heideltime [51] temporal tagger versus the previously used TIMEN [36] temporal tagger. (This choice was made because Heideltime supported more languages and an active development cycle.) Heideltime had no support for Portuguese (the primary language of Brazil) and the EMBERS software development team had extended Heideltime to support Portuguese by translating the underlying resources for Spanish to Portuguese. As it turns out, the simple translation of rules from Spanish to Portuguese was not sufficient and this affected the recall of one of the key models for Brazil, viz. the planned protest model. Since the planned protest model relies almost exclusively on the quality of information (specifically, date) extraction from text, its performance significantly deteriorated. This was subsequently corrected for the future by adding more rules and correcting existing rules (which were translated from Spanish) in Heideltime for Portuguese with the aid of language experts and extensive backtesting. Fig. 2.15 shows an example detection before and after the changes.

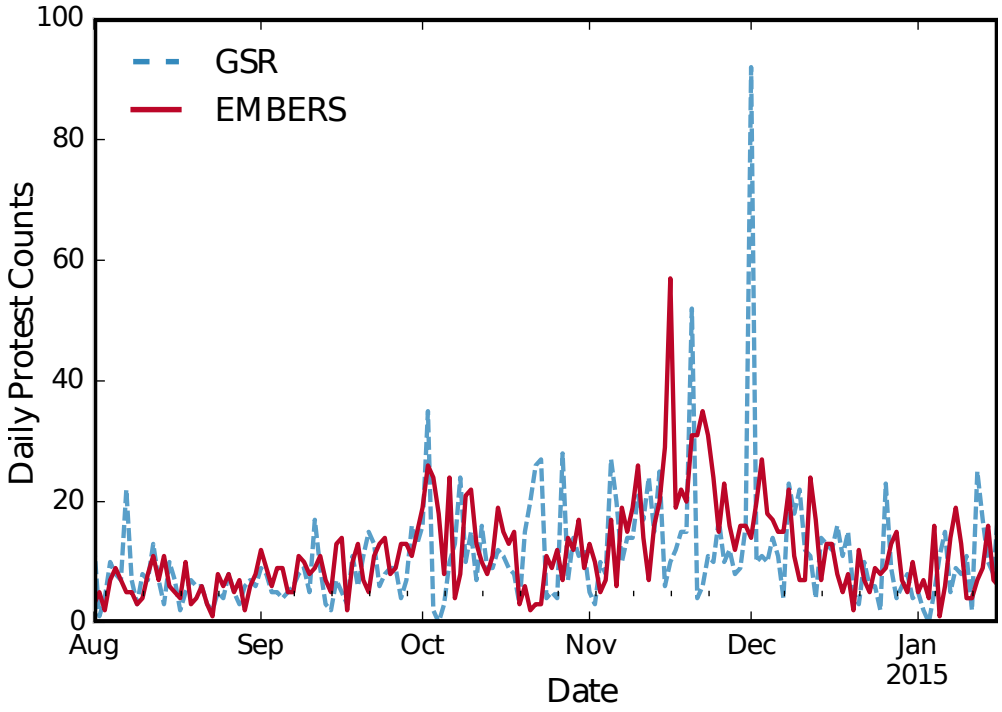


Figure 2.16: EMBERS performance during Mexico protests (Oct 2014).

Mexico protests (Dec 2014)

The days of December 2014 witnessed a continuation of the series of protests that began in October 2014 as described in Section 2.4.1. People turned out in huge numbers in different cities of Mexico demanding President Pena Nieto's ouster owing to the manner in which the case of the 43 missing students were handled. The protests were largely peaceful except for a few cases where vehicles were torched and windows and office equipment were broken.

EMBERS missed the huge single day spike on December 1st. Though having predicted a nationwide event for December 1, EMBERS failed to capture the individual cities where the protests would take place out and thus was unable to forecast the number of events accurately. On manual retrospective, it was found that the date, viz. December 1, was picked by the protesters due to its historical significance. This was the day when President Pena Nieto was sworn in (in 2012) amidst much controversy and opposition from many specific

Table 2.1: Comparison of performance measures under ablation testing. Social media sources contribute toward recall but, due to their noisy nature, lower other measures of performance.

Data Source	Quality-Score	Lead-time	Precision	Recall
Removing news and blogs	-16.48%	-55%	+35%	-14%
Removing social media	+8.42%	+30%	+79%	-33%

constituent groups. The manual analysis also led to the understanding of how special dates were mentioned by the twitterati **#1Dmx**. Dates mentioned using such abbreviations went unrecognized by the EMBERS system and was one of the main reasons for EMBERS not being able to capture the peak on December 1st despite its historical significance.

Brazilian Spring Onset

The EMBERS forecasts during the 2013 Brazilian Spring as shown in Fig. 2.6 was able to capture the peak but as can be seen the system was unable to capture the initial onset. See Section 2.7 for a detailed discussion of the limits of forecasting.

2.5 Ablation Testing

Different data sources provide different value to the forecasting enterprise. It is important that we understand the value of a data source w.r.t. its forecasting potential. In this section we describe ablation testing in EMBERS where the incremental value addition is evaluated for specific data sources. In particular, we are interested in determining the utility of using

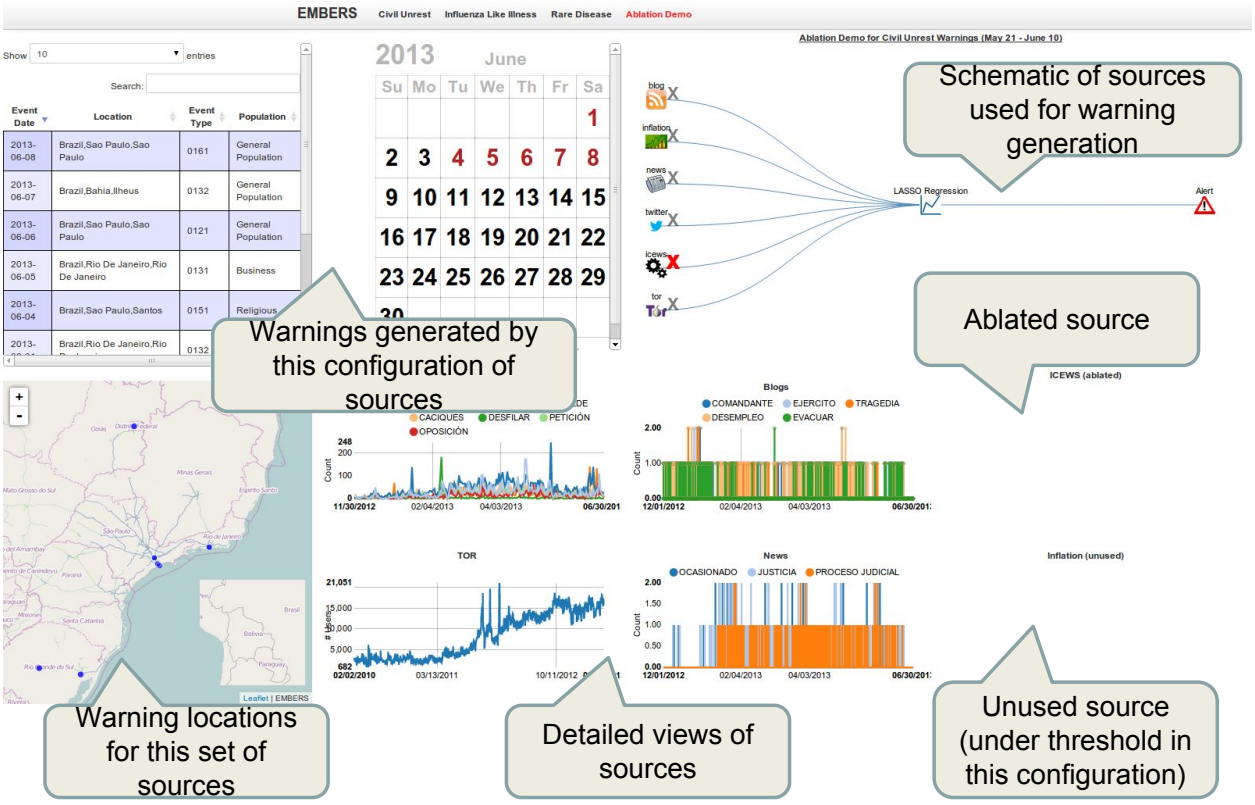


Figure 2.17: EMBERS ablation visualizer.

social media versus traditional media such as news and blogs. Table 2.1 shows the percentage improvement or degradation in performance measures when specific sources are removed. It clearly shows that social media sources are mainly necessary in achieving high recall but are not that useful in achieving high lead times, for which traditional media sources are required. This behavior is expected as social media is where daily chatter occurs whereas signs of organization and calls for protest often happen over (or are reported in) traditional meadia. Mainly, Table 2.1 makes it evident that to build a successful forecasting system we need a good mix of both traditional and social media sources. Fig. 2.17 shows a snapshot of the EMBERS ablation visualizer. The visualizer provides an analyst with the capability to selectively remove data sources and assess the differences in final alerts.

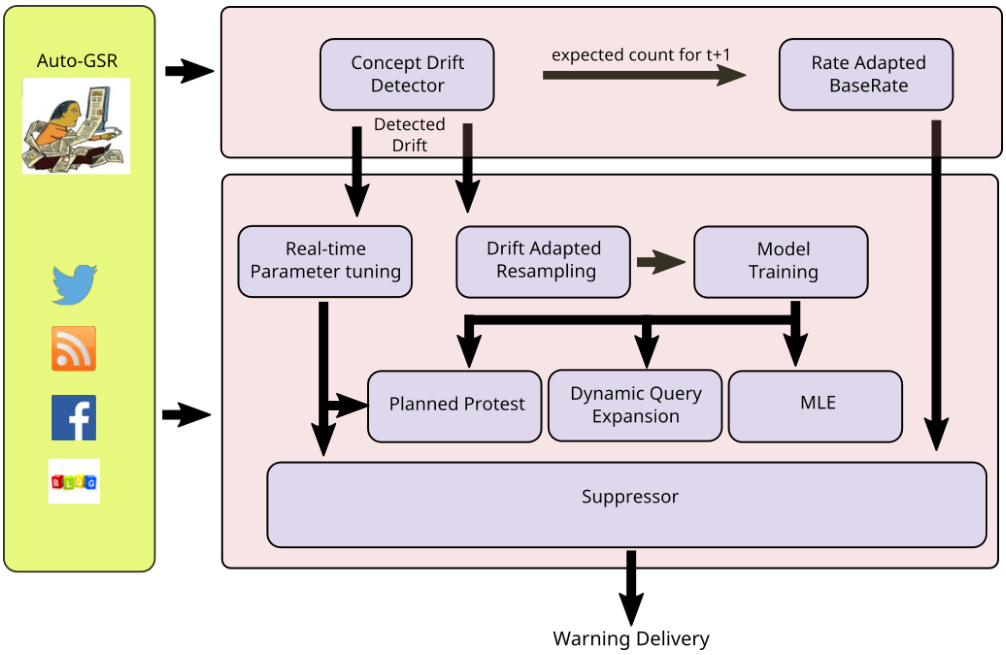


Figure 2.18: Drift Adaptation Framework

2.6 Model Drift

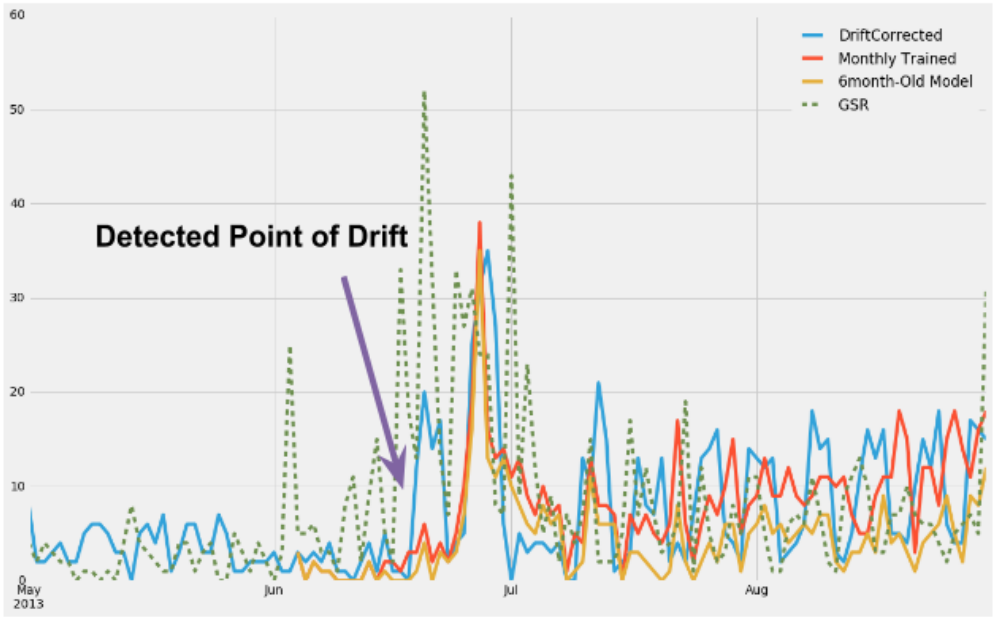
Model drift (also known as "concept drift" in the machine learning literature) is a critical issue given the "big data" approach espoused in EMBERS. Model drift occurs for many reasons, including a significant change in the event landscape that renders existing models to become outdated in their training. We make use of Prithwish et al. [9] "Hierarchical Quickest Change Detection" (HQCD) model to identify the time point after which the EMBERS civil unrest models drift away from the target (GSR). Once the point of change is detected, we then developed techniques to adapt the EMBERS models in real-time via automatic parameter tuning and model retraining to correct for the drift. The HQCD model detects changes in an online fashion across multiple sources viz. targets (GSR) and surrogates (twitter counts, news counts etc.).

After a changepoint is detected by HQCD, the forecasting models should be updated to adapt to the post-distributional changes to the target sources (GSR). We developed the following

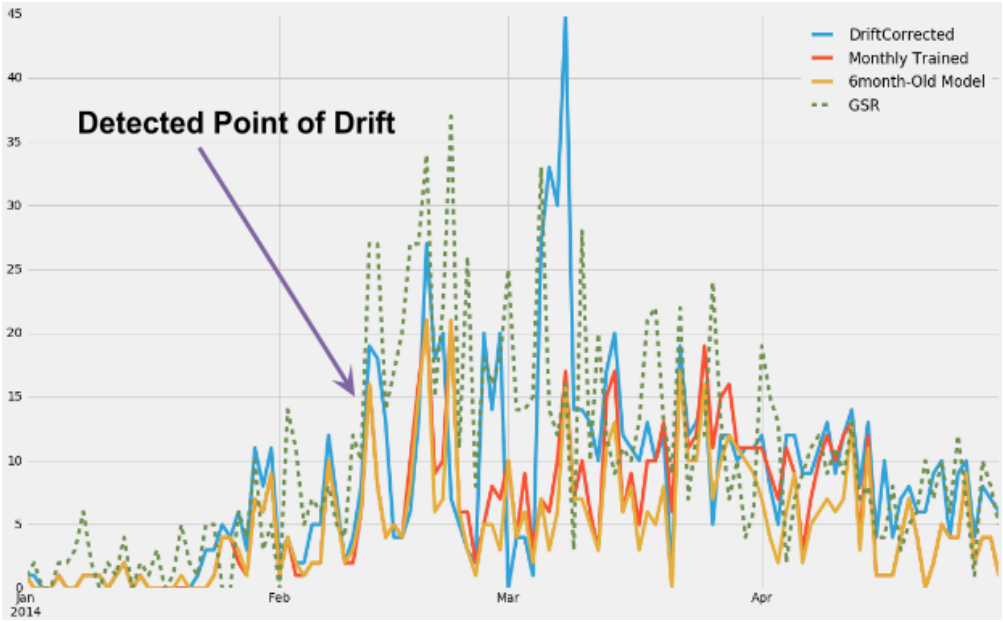
framework for model drift adaptation. The model drift adaptation involves two main steps – (1) real-time parameter tuning and (2) drift adapted re-sampling and model retraining. We also developed a rate adapted baserate model based on the expected counts output by HQCD. Real-time parameter tuning is done in order to allow models to adapt quickly to detected changes rather than wait for a long period to accumulate data from the post-change distribution for retraining. In real-time parameter tuning, different threshold parameters within models like the suppressor (predicted QS threshold), DQE (cluster warning threshold) etc., will be decreased or increased in small steps based on whether the post- change mean is higher/lower than the pre-change mean. Once we accumulate enough data from the post-change distribution, all forecasting models can then be retrained under the new distribution. In order to reduce the amount of time to wait for model re-training we devised a technique of drift adapted re-sampling to sample data for model retraining. In drift adapted re-sampling data is sampled from the pre- and post-change distributions based on the ratio of the pre- and post-change means of the target (GSR) as detected by HQCD. The overall model drift adaptation framework is shown in Figure. 2.18

The performance of the EMBERS models after drift adaptation was compared against (1) models diligently re-trained every month, (2) models trained only once every 6 months, and (3) the EMBERS delivered warnings. In Figure. 2.19 we observed that with model drift adaptation the counts of alerts per week (for Brazil) and per day (for Venezuela) aligns much closer to the counts of events.

Table. 2.2 shows the comparison of the drift adapted model against the comparative methods with respect to quality score, precision and recall. We show that with drift adaptation the EMBERS models show a much higher recall with minor sacrifice on precision. We also noticed that training regularly does not necessarily show much better performance than training once every 6 months.



(a) Brazilian Spring



(b) Venezuelan Spring

Figure 2.19: Weekly alert counts from various comparison models.The dotted green line is the ground-truth weekly counts from GSR. We can see that in both cases once the point of drift is detected with drift correction we are able to generate more alerts and the blue curve is able to match more closely the ground-truth curve than the rest.

	Events	Alerts	Matches	LS	DS	QS	Prob-M	LT	Precision	Recall
Venezuela										
Monthly Trained	793	512	474	0.9	0.94	3.69	0.89	5.01	0.93	0.6
6 months old	793	425	390	0.88	0.94	3.64	0.88	2.93	0.92	0.49
EMBERS-delivered	793	434	397	0.88	0.95	3.64	0.87	4.53	0.91	0.5
Drift Corrected	793	678	597	0.84	0.9	3.48	0.86	3.99	0.88	0.75
Brazil										
Monthly Trained	831	715	426	0.88	0.89	3.54	0.73	6.98	0.6	0.51
6 months old	831	408	304	0.86	0.87	3.46	0.77	3.46	0.75	0.37
EMBERS-delivered	831	581	425	0.87	0.88	3.49	0.77	5.89	0.73	0.51
Drift Corrected	831	756	500	0.89	0.89	3.55	0.76	5.27	0.66	0.6

Table 2.2: Quality score comparison of Drift corrected EMBERS alerts against 1) models diligently re-trained every month, and (2) trained once every 6-months and (3) the set of delivered EMBERS alerts.

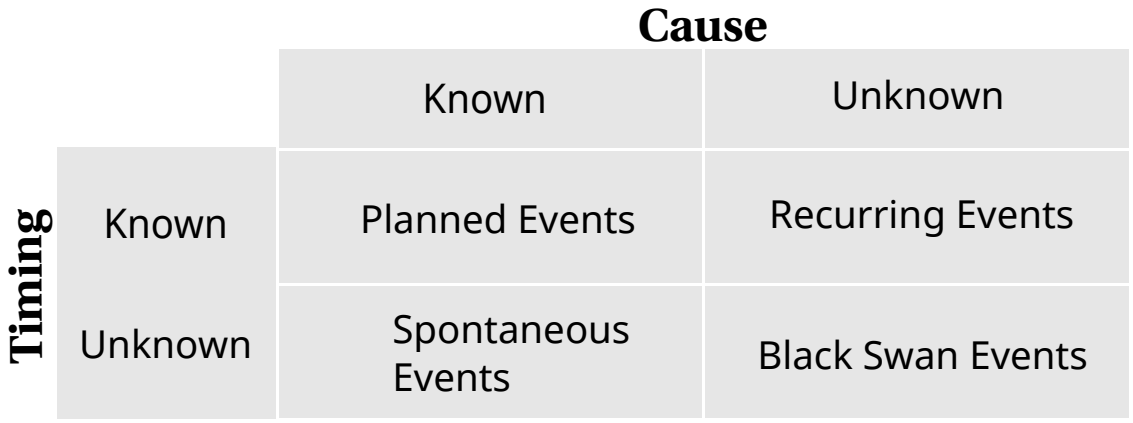


Figure 2.20: Studying the limitations of event forecasting.

2.7 Uncertainties in Forecasting

While examining the events of civil unrest closely in the past few years, it was clear to our team that events carry two distinct types of uncertainties: **cause** and **timing**. Fig. 2.20 summarizes these uncertainties.

Among all the incidents of civil unrest that we encounter, the largest and the most significant ones are planned events. These events are usually organized by political parties, labor and student unions. Since it takes a huge effort to organize protest demonstrations that attract

thousands, the organizers must disseminate information regarding the venue and the date and time. These announcements are posted on the organizers' websites and are widely shared on social media. By scouring our sources, it has been possible for EMBERS to accurately forecast the occurrence of these types of protests.

The recurring events take place on a regular basis. For instance, in Chile and Argentina the 'mothers of the disappeared' protest the disappearance of their children by the military dictatorships of the 1970s and 1980s on a certain particular day of the week and in the same plaza. In some countries with large Muslim populations, fighting and protests break out regularly after Friday evening prayer as people stream out of the mosques after listening to fiery sermons. These are typically small events but if they are reported as part of our GSR, EMBERS models will be able to forecast them.

The protests for which the causes are known but not the timing are staged spontaneously. These events are the outcomes of longstanding frustration and anger which fuel widespread protests in response to trigger events. Thus, the viral videos of police brutality or a sudden change in government policy can start a prairie fire of protests. The Brazilian Spring with origins in bus fares and which channeled public anger against corruption and government mismanagement is a classical example. The challenge here is not just to be aware of the underlying tensions that might erupt when an event occurs, but to also distinguish between events that do and do not perform as triggers. Algorithms to better model precursors is an area of further research that will aid further forecasting this class of events.

Finally, *black swan* events [52] are rare and truly unforeseen and can happen as a result of natural disasters, the sudden death of a leader, or even the sudden rise of a small group that can truly destabilize a nation. (Each of these types of events can in turn spark civil unrest.) For instance the rise of the Islamic State in Iraq and Syria (ISIS) has truly confounded policymakers all over the world. While there were other Sunni groups, from al-Qaeda to

al-Nusra, that contributed to instability, the rapid ascendance of ISIS, which did not depend on an isolated terrorist attack and burst out with a clear holding of territories as a full-scale insurgency, surprised most observers. It might not be feasible to forecast the beginnings of such events; however, once such movements have been initiated, models should be able to detect and forecast their momentum.

Chapter 3

Uncertainty-Aware Multi-Instance Learning for Event detection

In this chapter we will discuss about text classification approaches for event detection that are able to abstain when it is uncertain.In order to support real-word decision making, classifiers must not only be accurate but also be able to quantify uncertainty about a data-point. For example, most existing classifiers when provided with an out-of-distribution data-point (like images of numbers when the classifiers has been trained with images of cats) will likely classify them into one of the known classes and often times will do so with high certainty. The cost of such high confidence mistakes can be extremely high in applications like self-driving, healthcare etc. Ideally we want the classifier to say "I don't know" or yield a low confidence classification when presented with difficult and out-of-distribution instances. We propose a evidence theory based approach for uncertainty estimation in Multiple Instance learning settings. The proposed is capable of identifying which sentences causes uncertainty in forming opinions and thereby providing an interpretable way to understand text classification. The proposed model UQMIL is shown to outperform existing state-of-the-art methods

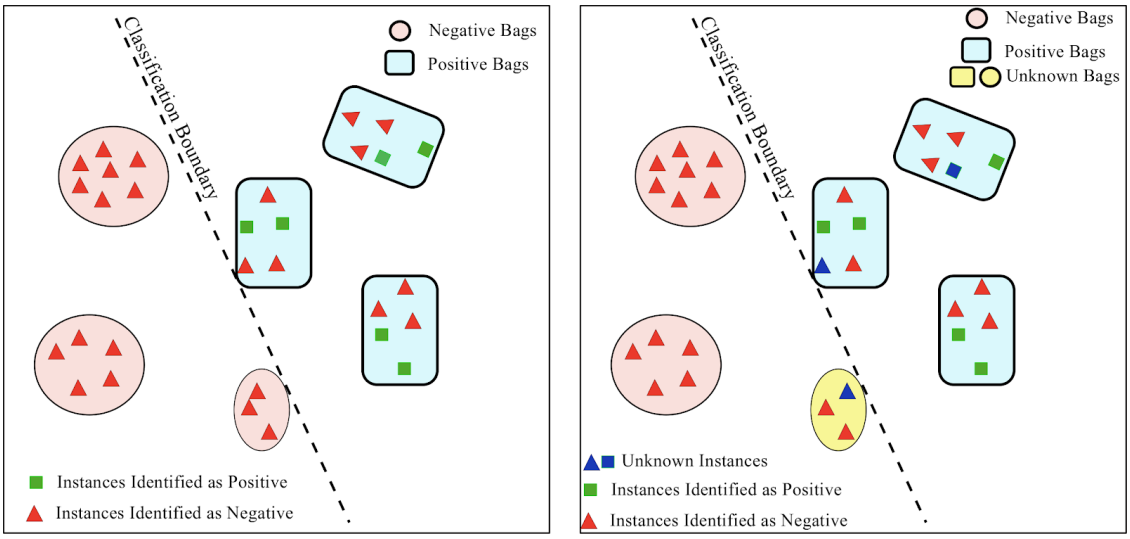


Figure 3.1: Comparison of a normal Multi-Instance learning model versus the proposed uncertainty-aware MIL model

under different data coverage settings.

3.1 Introduction

Classification is an important research problem in Machine learning. Often in recent times, there is a need for Machine learning models to work in tandem with humans.The expert judgement is reserved to humans. Current machine learning models often are trained without taking into account the amount of humans workforce available to the same problem, i.e., they are trained assuming there will be no human resources available to complete the problem. This leaves the ML model to be under-optimized. There is a lot of recent research that discusses learning models that has better uncertainty estimation capabilities.

Previous works like [50, 18, 35] provide methods for uncertainty aware classification. In applications like text classification, such methods, though capable of identifying a document as uncertain, are not able to pin-point which phrase/sentence was responsible for the uncer-

tainty. An understanding of which sentence/phrase causes perplexity can help identify the kinds of text/documents where more supervision is required. The paradigm of Multi-instance learning is a way to identify fine-grained / instance level labels from only coarse-grained supervision. We propose to extend the Multi-instance learning paradigm to KWIK (Knows what it knows) [34] settings to be able to ascertain which parts of a document cause uncertainty. Figure. 3.1 shows how we define an uncertainty-aware multi instance learning model. Under normal situations, we only have certain and uncertain instances. However, when expanding to MIL settings, we can also have certain bags with uncertain instances. A bag of instances can be labelled uncertain even if one of the instances in the bag is uncertain while the rest are negative. On the other hand, a bag is labelled positive if at-least one of the instances is positive. Note, in a positive bag, we can still have uncertain instances.

3.2 Related Work

Approaches for uncertainty estimation and calibration for classification tasks can be broadly categorized into the following: **I don't know classification** (IDK): Several methods for performing IDK classification like the Deep Abstaining Classifier (DAC) [53], Evidential Deep learning (EDL) [50] learn a $K + 1$th class for K class problem. Ziyin et al. [61] use portfolio theory to estimate IDK probability. Geiffman et al. [18] provide an approach optimized for a given data coverage setting. Often times in IDK classification the problem lies in selecting the best threshold for abstention.Alexander et al. [1] showcases an approach for threshold selection for abstention using curve optimization.

Ensemble Methods: The most popular approaches for estimating the predictive distribution variance is to use an ensemble method. Gal et al. [16] introduce different types of uncertainty associated with deep learning models and also illustrate that performing dropout

at test time is equivalent to a Gaussian process model. Traditional methods like boosting and bagging can also be used.

Max Margin classifiers: Max-margin classifiers [15] try to maximize the classifier margin with an aim towards better uncertainty mitigation. Xuchao et al. [59] uses metric learning to increase the distance between estimated embeddings of instances belonging to different classes and show this to improve classifier abstention performance.

External uncertainty estimation and calibration: Platt scaling [42] is used to calibrate the confidence predictions from any classifier. Kim et al. [26] introduce a trust score to identify when a model is uncertain about its prediction.

3.3 Preliminaries

In a general binary classification task we are presented with sample dataset $X =< x_1, x_2, ..., x_n >$, where $|X| = n$, and their corresponding labels $Y =< y_1, y_2, ..., y_n >$ where $y_i \in \{0, 1\}$. The label y_i is assumed to be sampled from a latent Bernoulli distribution paramterized by $p_i = P(y_i = 1|\theta)$. Most classifiers estimate the Bernoulli probability p_i directly from data. For example, neural networks estimate p_i using the softmax function. The uncertainty associated with the estimated p_i can be quantified using measures such as the information theoretic entropy $(-p_i \log p_i)$ or variance of the Bernoulli distribution $(p(1-p))$. However this assumes that the predicted probability is accurate. Gal et al. [16] show that neural networks can produce high confidence predictions even for a data point x^* that is far-off from training data distribution. Such unjustified high confidence predictions will lead to low entropy or Bernoulli variance. This showcases the limitations of vanilla models that directly tries to get a point estimate of p_i in understanding model uncertainty.

Vanilla binary classification models based on softmax function do not have a way to say a sample 'x' is out-of-distribution or noisy. In order to be able to do this, we make use of Dempster–Shafer Theory of Evidence [10] in this work. It is a generalization of the bayesian theory of subjective logic. Dempster-Shafer theory starts by assigning subjective probabilities to a set of possibilities. The superset of all such possibilities form the frame of discernment. Belief over a hypothesis represented by a subset of the frame of discernment can be obtained by summing the individual masses of all its subsets. Belief represents the amount of evidence in favor of a proposition q (within the hypothesis). The absense of belief, that is $1 - belief$, represents the *plausibility* of proposition q. It is an upper bound on the possibility that the hypothesis could be true.

Subjective Logic [27] formalizes the notion of evidence to Dirichlet distribution (or beta distribution when only number of categories $K = 2$). Given belief b_0 and b_1 for the two categories of our beta distribution, the overall belief can be quantified as follows:

$$b_0 + b_1 + u = 1 \tag{3.1}$$

Here, $b_0, b_1, u \geq 0$. u represents the uncertainty mass. If e_k is the evidence associated with belief b_k and assuming uniform prior, we have:

$$\begin{gathered} \alpha = e_0 + 1, \beta = e_1 + 1, S = \alpha + \beta \\ u = \frac{K}{S}, b_k = \frac{e_k}{S} \end{gathered} \tag{3.2}$$

Here, S is overall dirichlet strength (or beta strength). α and β are the beta distribution parameters. From Equation. 3.2, we can see that higher the evidence e_k, lower the uncertainty. Finally, we know beta distribution is the conjugate prior of the categorical distribution. Assuming the prior for p_i is a beta distribution parameterized by $Beta(\theta|\alpha, \beta)$, we have

$p_i = P(y_i = 1|\theta) * P(\theta|\alpha, \beta)$. Marginalizing over θ we have $p_i = \alpha/(\alpha + \beta)$, the expectation of beta distribution. Now if we design a model to predict the amount of evidence e_k instead of directly estimating the probability p_i one should be able to quantify uncertainty more efficiently as per Equation 3.2 than a model that produces a point estimate of p_i. Such a model can be optimized by minimizing the overall loss after marginalizing out the latent θ variable. Thus if we use the log-loss (or cross entropy) the overall loss functions looks like:

$$\mathcal{L}(\phi) = \mathbb{E}_{p(\theta|\alpha,\beta)}(-y_i \log \theta - (1 - y_i) \log(1 - \theta)) \tag{3.3}$$

From Sensoy et al. [50], we can get reduced form of the above equation after marginalizing out θ as

$$\mathcal{L}(\phi) = \sum_{j=1}^{K} y_{ij}(\psi(S_i) + \psi(\alpha_{ij})) \tag{3.4}$$

where $K = 2$ the number of categories and ψ digamma function.

3.4 Problem Definition

In the previous section we saw how we can use the theory of evidence to quantify uncertainty in a binary classification task. In this section we will extend the formulation for use in Multiple Instance Learning (MIL) [11] settings. In MIL, each data sample x_i, called a bag, is comprised of at most n instances x_{ij}, whereas the label y_i is only available at the bag level. The objective is to build a machine learning model that can efficiently classify bags x_i and at the same time identify infer the labels for each instance x_{ij}. In event detection paradigm, a bag is represented by a document and an instance by the individual sentences within the

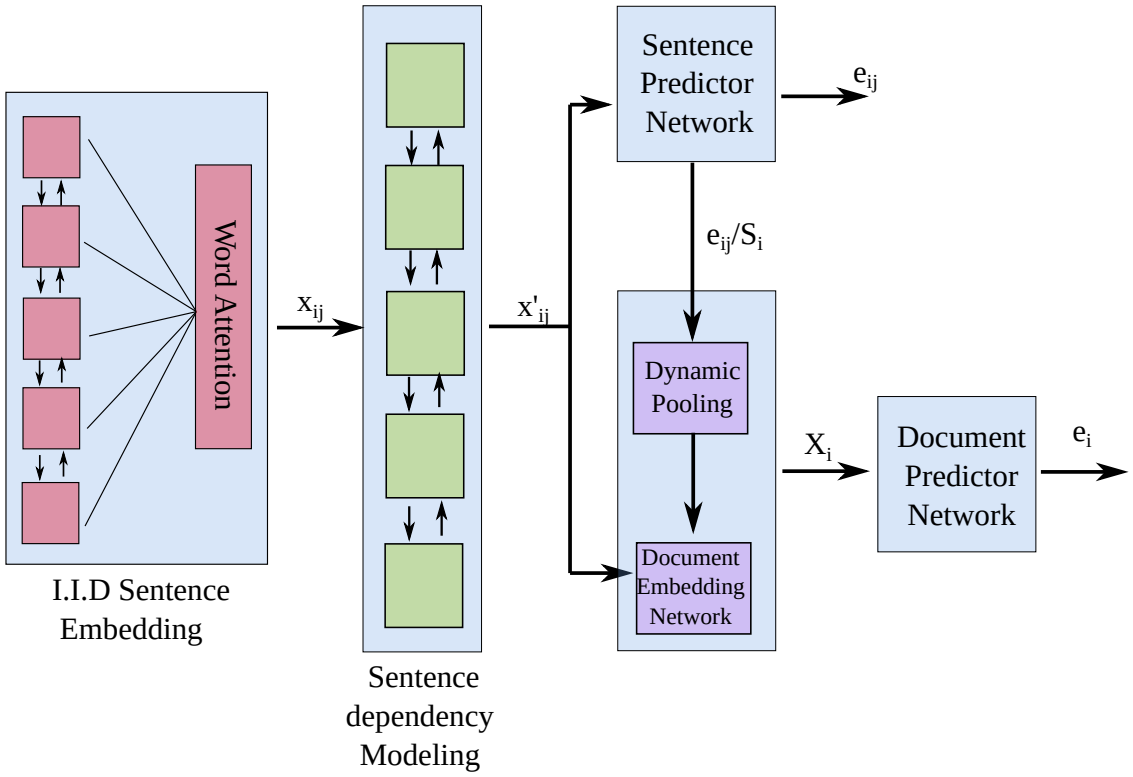


Figure 3.2: UQMIL Model Architecture

document. In our problem setup, we extend Evidential Deep Learning [50] to MIL settings, wherein we will design an Deep Learning model to estimate instance level evidence e_{ij} and thereby the instance level uncertainty u_{ij} and use it to estimate overall bag level evidence e_i and uncertainty u_i. Next we will describe our proposed model UQMIL for instance level uncertainty estimation.

3.5 Proposed Model

We propose a MIL framework for instance or sentence level uncertainty quantification using the base architecture of Hierarchical Attention Networks(HAN) [57]. Figure. 3.2 showcases the proposed model architecture. Initially, for an input document x_i, the model processes

each individual sentence x_{ij} independently using a bi-directional GRU, followed by an dot-product attention mechanism to obtain the sentence level embedding m^s_{ij}. Thus,

$$
\begin{aligned}
H &= \text{Bi-GRU}(\text{EMB}(x_{ij})), && H \in \mathcal{R}^{t*d} \\
U &= tanh(HW + b_u), && W \in \mathcal{R}^{d*d} \\
a_t &= \frac{exp(u_{ijt}^T u_w)}{\sum_t exp(u_{ijt}^T u_w)} \\
m^s_{ij} &= \sum_t a_t * u_t
\end{aligned}
$$

Here a_t is the attention weights, and u_w is the word attention context vector. The sentence level embeddings m^s_{ij} are obtained independent of other sentences present in the document. However, in a document setting, i.i.d assumptions of instances may not be valid. The dependency between sentences is modeled using another bidirectional GRU. The output of the sentence-level GRU goes through a non-linear transform before being passed to the Sentence Predictor Network (SPN). The SPN produces the sentence level evidences e_{ij0} and e_{ij1}. Finally, in order to estimate the bag level embedding m^d_i we use weighted average pooling with the positive class belief of a sentence as the weight. Note $b_{ij1} = e_{ij1}/(e_{ij1} + e_{ij0} + 2)$. Document embedding m^d_i is then passed through a linear layer, followed by an exponential transform to obtain the document evidences e_i0 and e_i1. We term this model UQMIL.

We can also perform a dynamic aggregation of sentence embeddings to get the document embeddings. The dynamic pooling network comprises of an attention network similar to the word level attention which takes as input the predicted sentence level evidences. We term this variant of our proposed framework as UQMIL_dynPool.

The overall model is optimized using the loss function specified by Equation. 3.4. The loss function is applied at both document and sentence level, with bag level label propagated to the instance level. Additionally, in order to prevent the model from assigning evidence to incorrect class we add an additional KL-divergence loss that penalizes the model from producing output close to the "I don't know" state of maximum uncertainty. The evidence regularization loss is given by

$$\mathcal{L}_{KL} = \lambda_t * KL(B(\theta|\hat{\alpha}, \hat{\beta})||B(\theta|| < 1, 1 >)) \tag{3.5}$$

wherein $\hat{\alpha} = 1$ if $y = 1$ else $\hat{\beta} = 1$. λ_t is an annealing coefficient that increases linearly with number of epochs elapsed. It is present to prevent the model from reducing to the maximum uncertainty case initially. Thus the model is trained to optimized the following overall loss:

$$\mathcal{L} = \sum_i \left(\mathcal{L}^d_{KL} + \mathcal{L}^d(\phi) + \sum_j ((\mathcal{L}^s_{KL} + \mathcal{L}^s(\phi))) \right) \tag{3.6}$$

wherein $\mathcal{L}^d$ represents document level loss and $\mathcal{L}^s$ the sentence level loss.

In the following section, we detail the experimental settings for the evaluation of our model.

3.6 Experiments

3.6.1 Dataset

Our dataset comprises of manually labeled Military Action and Non-State Actors (MANSA) English language articles obtained from national news papers of Bahrain, Egypt, Iraq, Jordan, Lebanon, Qatar, Saudi Arabia and Syria. The data period ranges from 2016-08 to 2018-09. The dataset consists of 128620 articles with about 33% of the documents labelled

as MANSA event article. We split the dataset temporally to create train, test and validation sets. Data upto 2018-05 is used for training and articles from the month of 2018-06 is used validation. The articles from the remaining months is used for testing.

3.6.2 Experiment Protocol

We anonymize all entities such as locations, organizations and person names during pre-processing to prevent the model from being biased towards a particular entity. For all comparative methods we use the same network architecture based on Hierarchical Attention Networks.All models are trained for a total of 5 epochs with a batch size of 32 and using the ADAM optimizer.

3.6.3 Comparative Methods

We compare our method against the following state-of-the-art models:

- **SelectiveNet** [18] optimizes selection prediction for a given coverage value using interior point optimization method. Ideally, it is required we train the SelectiveNet model separately for each coverage configuration. However this is expensive and is unfair the rest of the models. We find setting $C = 0.8$ works best across all coverage ratios.

- **Deep Abstaining Classifier** [53] adapts the traditional cross entropy loss to learn a $K + 1$th abstention class.

- **Metric Learning** [59] proposed by Xuchao et al., learns to maximize the distance between document embeddings of different classes. Additionally it uses McDropout [16] to estimate the final prediction variance. However we only make use of the softmax

response of the final layer as the confidence estimate and do not use McDropout as it is test time heavy and is sensitive to the number of samples obtained.

- **Softmax Response** [17] refers to the maximum estimated probability of the base classifier.
- **Evidential Deep Learning**(EDL) [50] learns to model beta distribution counts instead of directly estimating the Bernoulli probability.
- **Platt Scaling** [42] refers to the external calibration performed over estimated model probabilities as shown by Niculescu et al.

We do not compare against McDropout [16] and other sampling based approaches as they are test time heavy and their performance is sensitive to the number of samples drawn.

3.6.4 Metrics

We evaluate the models in terms of the following metrics:

1)**F-1** score at different data coverage ratios

2)**Expected Calibration Error**(ECE) is the sum of absolute differences of the accuracy and mean-confidence under each bin corresponding to a given confidence range.

3.7 Results and Discussion

In this section we showcase the performance of our proposed model UQMIL. First we evaluate the classification accuracy of our model for event detection in comparison with baseline methods (in section above) under multiple data coverage configurations. Next we

	100%	95%	90%	85%	80%	75%
Softmax Response	0.640	0.668	0.685	0.712	0.730	0.721
MetricLearning	0.638	0.664	0.686	0.704	0.722	**0.740**
DAC	0.646	0.668	**0.692**	0.705	0.714	0.737
SelectiveNet	0.628	0.635	0.647	0.655	0.646	0.555
EDL	0.614	0.639	0.655	0.672	0.688	0.680
Platt Scaling	0.611	0.628	0.655	0.673	0.688	0.657
UQMIL	**0.653**	**0.674**	**0.697**	**0.720**	**0.747**	**0.749**
UQMIL_dynPool	**0.656**	**0.678**	**0.697**	0.718	0.728	0.721

Table 3.1: Performance comparison of F-1 score for event detection under different data coverage ratios. Here 100% coverage means the classifier is run on the whole dataset, while a coverage of 90% means the classifier performance is only evaluated on 90% of the data (ordered by confidence). The remaining 10% of data, comprising of low confidence (or high uncertainty) data points, is set aside for human intervention.

analyze if the estimated bag-level probabilities for the event class are well calibrated. Finally, we qualitatively analyze the sentence (or instance) level uncertainty quantification.

3.7.1 How well does UQMIL perform under different data coverage configurations?

Table. 3.1 shows the performance of our proposed model UQMIL and UQMIL_dynPool compared against that of existing state-of-the-art methods mentioned in section 3.6.3. From the table we can see that the proposed model UQMILoutperforms all existing methods in all coverage ratio settings. UQMIL_dynPoolalso performs equally well and also beats almost all comparative methods. The Metric learning based approach and DAC provide the second best performance. Softmax-Response also showcases similar performance as DAC and Metric learning. The poor performance of Platt scaling is possibly because of the underlying model softmax responses are well calibrated(as softmax response based approach showcases better F-1 scores).

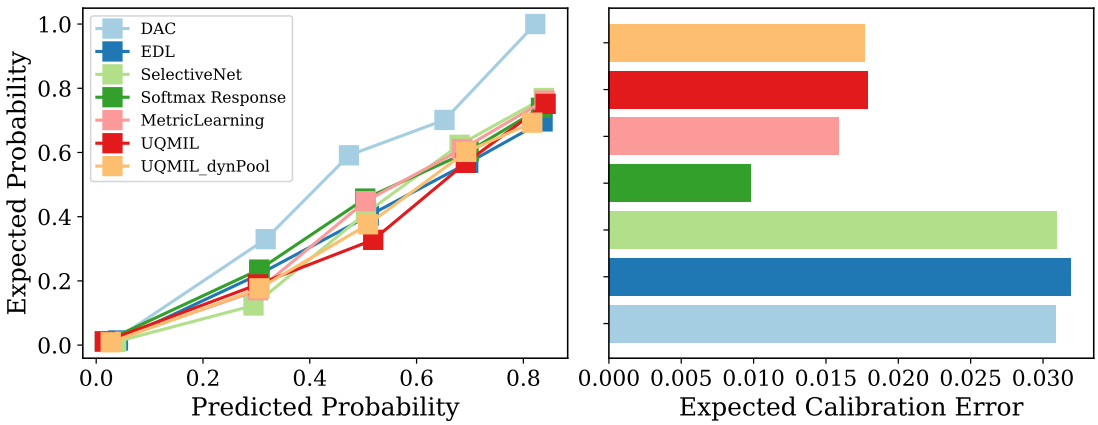


Figure 3.3: Evaluation of model calibration in terms of Expected Calibration Error (ECE). Softmax-Response achieves the best calibration error as it directly optimizes the log-loss. UQMIL , UQMIL_dynPool and the metric learning based approach perform equally well while DAC and EDL showcase poor calibration of predicted probabilities.

3.7.2 How well are the model probabilities calibrated?

Figure. 3.3 showcases the reliability curve of the various comparative methods. We can see that softmax response showcases the best calibration. This can be because it directly optimizes the default cross-entropy loss using a softmax function which is known to provide good calibrated probabilities. The calibration of both proposed models UQMIL , UQMIL_dynPool and Metric learning are similar, while EDL and DAC perform the worst.

3.7.3 How good is the uncertainty quantification at the sentence (or instance) level?

Table. 3.2 showcases some examples of sentences identified as unknown or "I don't know". From the table we can see that the model is able to identify the presence of non-english text in English language articles as "I don't know". Similarly it identifies examples of arrests by state actors as unknown. Arrest events are not covered under our dataset. However our

U	label	Sentence
0.51	0	'The Popular Forces forces were able to dismantle a terrorist cell that was planning to target religious processions in Muqdadiya during the first Ashura'
0.432	0	'A security source in Babylon told the correspondent of NINA that the force of the popular crowd in Al - Farisiya region of Jurf al - Sakhr , arrested an armed man carrying explosives , tried to sneak into one of the sites belonging to the popular crowd',
0.436	0	The combat force of the 24th Brigade in the popular crowd carried out the first Ashura security operation to track the cells in the districts of Babil and Taqul north of Muqdadiya district
0.407	0	Kurdish Self - Management forces (Kurdish Democratic Union Party - PKK branch) arrested 10 civilians from their place in February 23 Street in the middle of Raqqa city and took them to unknown place , on September 6 , 2018
0.383	1	**ARABIC TEXT**
0.413	0	'2 children killed on July 31 , 2018 , due to explosion of a landmine planted by ISIS in Ashayer village of al Boukamal city in Deir Ez - Zour governorate eastern suburbs , before retreating from it'

Table 3.2: Some examples of sentences identified as uncertain (or "I dont Know" (IDK) instances). From the examples we can see that some of the IDK sentences are possibly data points with label noise, different language text and instances talking about arrests.

dataset covers hostage taking subtype of events. Arrest events can be considered as Hostage taking, but led by state actors which possibly is the reason for the uncertainty associated with the classifier output evidences. Thus overall we can see UQMILis able to extract meaningful sentences as uncertain or unknown sentences. This will help prioritize human input.

3.8 Discussion

In this chapter, we present a novel way for extending evidence theory to Multiple Instance Learning (framework). By extending evidence theory to MIL framework, we are able to obtain fine grained sentence level uncertainty quantification. Such fine-grained uncertainty will help prioritize human input when available. Through extensive evaluation and analysis we illustrate the strong performance of our model.Additionally we provide examples of sentences classified "I dont Know" (IDK) for a qualitative analysis. In the future, we plan to extend our model to an active learning framework wherein human feedback for IDK sentences can be used to improve overall model performance.

Chapter 4

Hybrid Micro Tasking Framework for Event Coding

In the previous chapter we talked about techniques for fine-grained uncertainty quantification in event detection. In this chapter we will talk about how to extract relevant information about 'who, when, where and why' of an event from articles output by such an event detection pipeline. Existing event coding systems in practice that churn out real-time "event data" are generally of two types – 1) fully automated and 2) human annotated. Human annotated or Hand-coded systems like SPEED [7] and EMBERS AutoGSR [47] are generally considered highly reliable but lack scalability and are highy expensive to run for a prolonged period of time. On the other hand automated event coding systems like ICEWS [5] and TABARI [48] though highly scalable as compared human coded systems still have several limitations. (i)First, only few sentences of a document are scanned and used for extraction purposes. This could lead to a lot of context surrounding the event to be missed and cause increased duplication of events by not recognizing re-reporting of historic events.(ii) Secondly, event extraction is still performed based on hand coded dictionaries and rules. The usage of such

dictionary based pattern matching specifically in GDELT [32] leads to increase in the number of false positives and thereby affects overall reliability. (iii) GDELT [32] and ICEWS [5], though the latter makes use of machine learning, are still lagging behind in leveraging some of the significant improvements made in natural language processing in recent times. Also, Wang et al. [54] found such automated systems to have very low correlation with comparable manually curated ground-truth, thereby showcasing low reliability. GDELT matched SPEED events (events happening on the same day) about 17.2% where as ICEWS agreed on 10.3% of events. Wang et al. [54] also found only 49% of events extracted by GDELT to be valid i.e., refer to an actual protest event. ICEWS was identified to be more robust in terms of validity but still was vulnerable to duplicate events. The existing systems are trained on a specific taxonomy called CAMEO and will need significant manual effort and re-design to adapt to new domains. Finally all existing event coding systems work primarily with English text and use translation when trying to code local languages (as focused here).

In this problem, we try to alleviate some of the above mentioned issues by building an event coding framework that understands what it doesn't know and accordingly sends uncertain instances for human intervention. With such a system we will be capable of reaching the reliability and validity of human coded systems while at the same time be able to scale to hundreds of thousands of articles like the automated systems. Also one will have the choice to sacrifice recall (and scalability) for precision (and reliability) or the vice versa depending on our use case and domain of interest.

4.1 Related Work

Three categories of related work are briefly discussed here:

Semi-Automatic coding systems like SPEED (Social, Political, and Economic Event

Database) [7] and EMBERS AutoGSR [47] use a combination human and automated techniques for identifying events. Saraf et al. [47] show that by introducing automation to certain parts of the coding pipeline (like news collection, spam filtering, geo-coding) it is capable of reducing about 70% of time spent by humans on a article.

Automated Event coders include statistical as well as linguistic and lexicographic techniques. Schrodt et al. [48] introduced one of the earliest event coders TABARI (Textual Analysis by Augmented Replacement Instructions). It searches for hand coded patterns on only the first few sentences of a news article to encode an event of interest. TABARI was succeeded by JABARI and PETRARCH [49]. BBN's SERIF (statistical Entity and Relation Information Finder) is another state-of-the-art event coder using several NLP components to identify triplets of type actor-subject-target from article text (at both sentence and document level). It is to be noted that the SERIF encoder is not available for public use. The ICEWS (Integrated Crisis Early Warning System) [5] makes use of the SERIF encoder to parse hundreds of articles to produce a real-time database of events. GDELT(Global Database of Events, Language and Tone) [33] another event coding system that churns out events from a larger geographical area and categories makes use of an enhanced version of the TABARI parser. All of the above mentioned systems suffer from issues of duplication, reliability and validity apart from inability to easily port to new languages.

Semantic parsing is the process of converting a natural language text into a machine understandable formal meaning representation. Kamath et al. [28] provides a summary of the various techniques for semantic parsing. Most existing state-of-the-art semantic parsing techniques like Berant et al. [3, 4], Dong et al. [13], and Nguyen et al. [40] only work at the sentence level and are focused on extracting triplets of the form *relation(subject, object)*.Thus such methods are not directly applicable to event-coding, wherein required information might be spread across multiple sentences of a document.

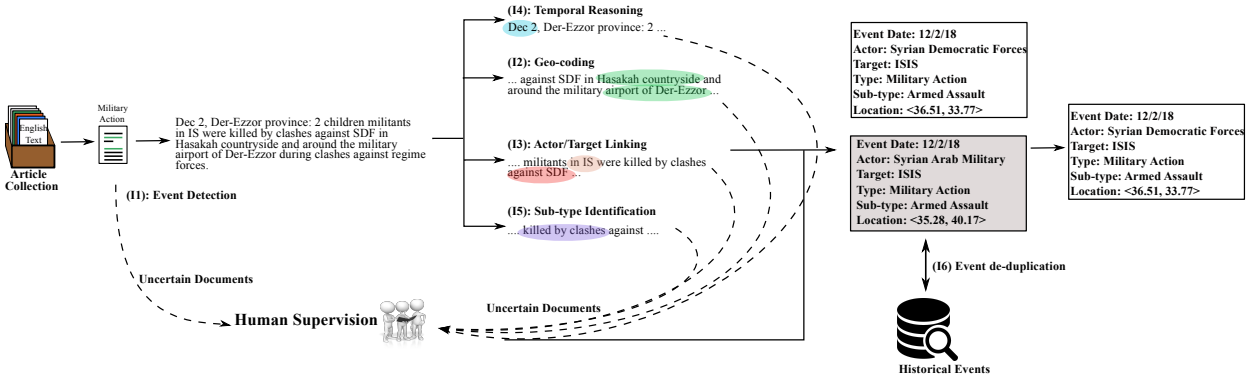


Figure 4.1: The proposed hybrid event coding framework

4.2 Problem Formulation

In our framework we break down the entire task of event encoding into multiple micro-tasks (or sub-tasks) and each such micro-task is built with uncertainty in mind. An illustration of our framework is shown in Figure. 4.1. As shown in the figure our system consists of the following micro-tasks

- **Event Detection** is the process of classifying if a news article is reporting an event of interest or not.
- **Geo-coding** is the task of identifying location names from text and grounding them to a latitude/longitude on earth.
- **Actor / Target Linking** involves identifying different entities mentioned in the article and linking them to known actors in our dictionary.
- **Temporal Reasoning** involves identifying the exact date in which an event took place by resolving all direct and relative dates mentioned in the text.
- **Sub-type Identification** involves identifying the event sub-type. This could be either the reason for the event (like Economic, Religious) in case of civil unrest or the kind

of event (like bombing, hostage taking).

- **Event De-duplication** refers to the task of identifying if the current article refers to an event that was already extracted and if it refers to an already extracted event, the current article will be discarded (assuming the article reports only the duplicate event).

Each one of the micro-tasks is explained in detail in the following sections.

4.3 Event Detection

The event detection part of our system is composed of a system of binary classifiers followed by a random forest ensemble to combine the predictions of individual classifiers. The classification models used includes methods like the Hierarchical Attention Networks (HAN), Support vector machines, Naive Bayes, Decision Trees, AdaBoost etc. In total we use a system of 16 methods and each such method is used to create two models - one with just the title of the article and another with the whole text of the article. Finally the predicted class probabilities by each model is used as features in a Random forest model to get the final prediction. The confidence/uncertainty score of prediction is taken to be the probability of the max class.

4.4 Geocoding

We built a supervised disambiguation strategy for disambiguating mentions of location in text. The geolocation extraction performance of this new strategy is compared with the state-of-the-art geolocation model Mordecai (https://github.com/openeventdata/mordecai).

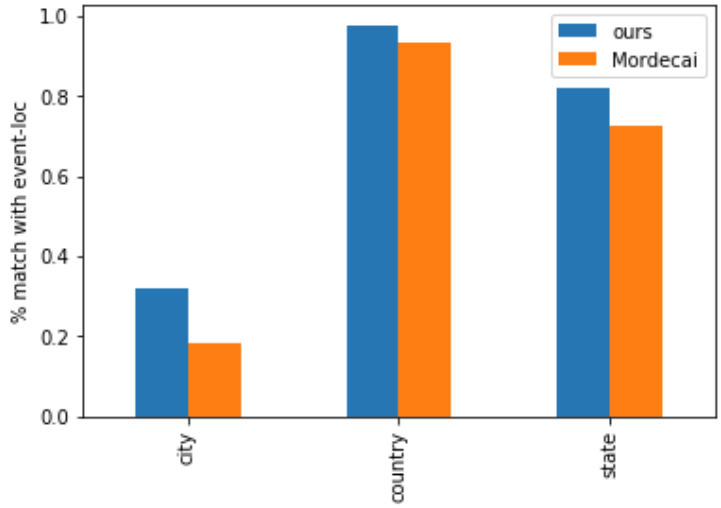


Figure 4.2: Performance comparison of the proposed geocoder vs state-of-the-art geo-coder Mordecai

Both models are evaluated on English articles for the month of 2018-05 (we use only English for comparison, as Mordecai is only available for English). There are in total 1,743 events, of which 449 events have an approximate location, i.e, the location might not be directly mentioned in the text. We check if the event-location is among the disambiguated locations obtained from the location expressions in the text. The location expressions are obtained using a Named Entity Recognizer. The performance is shown in Figure. 4.2 .

Our geocoding model works by first extracting all named entities namely Locations, Organizations and Person names, from text. We then query an elasticsearch database of Geonames for all possible expansions of a location entity mention. Once all possible expansions for a mention is obtained, we then decide which among the expansions is referred to in the text. This process is called Geo disambiguation. Most of the existing state-of-the-art methods like Kamalloo, et al. rely on hand-coded rules/hypotheses, e.g., location names mentioned consecutively share a common ancestor, names only refer to the most populous expansion, etc., for disambiguating location names. Unlike these methods, we try to learn the rules

automatically using machine learning. For each location expansion we built a feature set based on its frequency, the corresponding country and state frequency, distance to nearest mention of the corresponding country/state (both before and after), relative population of this expansion with respect to others etc. These features are then fed into a Random Forest classifier to decide which location a given named entity refers to. Since we do not use any features based on the words in the text, our method is language agnostic unlike other supervised geo disambiguation methods like Mordecai. Mordecai makes use of the words/context around a named entity mention to identify the country. We also query the organization names obtained to see if it mentions a country or state (for example, the entity "Indian Department of Treasury" gives hint about the country India).

4.5 Actor/Target Linking

In this section we identify how entities in text identified as subject/object are converted/linked to actors of interest for our purpose. Proper nouns or entities recognised as PERSON/ORGANIZATION during the language enrichment step are first checked against the list of known Actors and their aliases. We specifically make use of Jaccard similarity (in terms of words/terms instead of characters) to identify the best match. If an entity recognised by the NER system is not found in our list of known actors we call the actor as "Unknown" and also mark it for human supervision.

For text detected as subject or generic nouns, e.g., hospitals, military, etc., we make use of wordnet to identify their semantic category and use the identified semantic category as the final detected actor if it's in our generic actors list (generic actors list includes actors/targets like military, terrorists, buildings, people etc.).

4.6 Temporal Reasoning

In this step, we build techniques to understand which date the event occurred. For this, first we make use of Heideltime [51] temporal tagger for resolving any date/time mentions in the text. Heideltime supports innumerous languages and has exhaustive set of patterns for identifying mentions of relative temporal expressions and resolving them with respect to anchor date. The anchor date in our case is taken to be the article publication date if this information is available (in the meta tags) else it is assumed to be the date at which the article was crawled.

4.7 Sub-type Identification

For sub-type identification we make use of word-net and word-embedding based similarity measures with respect to the domain keywords.

4.8 Event De-duplication

Once an event is extracted from an article, we need to identify if the event refers to already extracted event in the database. If it refers to an already extracted event, then the current event is a duplicate and will be discarded. This entire process is called event de-duplication. In our framework, we perform de-duplication by comparing the source article text with the article text of all events within the last 2 days and deem it to be a threshold if the article similarity is higher 0.8. The article similarity is calculated as a weighted sum between 1) entity similarity, 2)location similarity and 3) text similarity. Text similarity is calculated using cosine similarity. While entity and location similarity are calculated using Jaccards

metric [23].

4.9 Experimental Settings and Evaluation

All models, unless otherwise stated, are trained on data up to May 2018. Data from June 2018 is used for validation. The results are reported on the remaining months, i.e., August through October 2018. Note, Nov-Jan 2019 is not used for evaluation as these months only have ground-truth for Military Action events.

As stated in the previous sections, the machine learning models perform five micro-tasks in total - Event Detection, location identification, actor/target extraction, event-subtype identification, date extraction. For all micro-tasks, except for date extraction, the ML models can abstain from making a prediction. In such cases the document is sent for human supervision and the annotators are asked to provide an answer for that micro-task. Note the annotators only fill-in the information for the asked field (as defined by the micro-task) and do not change any other fields extracted by the ML models.

We report the performance of our overall system using the Performance metrics as defined by the IARPA OSI Mercury program. The performance metrics include -

- **Event Detection Performance Metrics**
  - Precision/Recall/F1 metrics: These metrics are defined as expected and are standard.
  - Risk-Coverage Curve: Risk is defined as the percentage of false classifications (False Positives + False Negatives) w.r.t to the number of data points that are considered / covered. Coverage is defined as the percentage of data remaining after churning out documents on which the model score is less than the identified

threshold. The curve is plotted by first sorting the model predictions in terms of its confidence score and moving the threshold from 0 to 100th percentile value of the confidence score.

- **Event Encoding Performance Metrics**
  - **Quality Score:** A score out of 4 and is a weighted sum over the extraction performance for each individual field in an event record. The performance for each field is a score between 0 and 1, with 1 indicating perfect extraction and 0 referring to fully wrong extraction. The components of QS are
    1. Actor score
    2. Target Target score
    3. Target Status score
    4. Location score
    5. Event Sub-type score
    6. Date score
  - **Macro and Micro averages:** The macro and micro averages of the above mentioned metrics are defined similar to the macro and micro averages in classification problems. For Macro-average, we first calculate QS, Precision, Recall and F-1 per document (i.e. the extracted events in a document are only evaluated against the ground-truth events in that document) and then take the average of the scores per document. For Micro-average metrics, we evaluate all extracted events against all ground-truth events irrespective of which document they came from.

4.10 Results and Discussion

In this section we will present the scores obtained by our event encoding system. Specifically we aim to answer the following major questions.

1. What is the event detection performance of our AutoGSR system?
2. How good is the uncertainty/confidence characterization of the event detection model?
3. What is the performance of human annotators for each micro-task?
4. How many documents are abstained per micro-task?
5. What is the event extraction performance of the ML only System?
6. What is the event extraction performance of the Hybrid System?
7. What is the overall conclusion for performing MANSA event encoding?

In the following subsections we will look at each question individually.

4.10.1 What is the event detection performance of the ML-only system?

Here we present the performance of our ML system for the micro-task of event detection. The event detection micro-task involves predicting if an article contains an event or not.

The tables Tab. 4.1 and 4.2 provides the performance metrics our event detection system for English (Tab. 4.1) and Arabic documents (Tab. 4.2) for the test period.

	Precision	Recall	F1-score	Support
No-event	0.97	0.95	0.96	15843
Event	0.75	0.84	0.79	2760
micro avg	0.93	0.93	0.93	18603
macro avg	0.86	0.90	0.88	18603
weighted avg	0.94	0.93	0.94	18603

Table 4.1: Event Detection performance for English documents in test set.

	Precision	Recall	F1-score	Support
No-event	1.0	1.00	1.00	28643
Event	1.0	0.36	0.53	73
micro avg	1.0	1.00	1.00	28716
macro avg	1.0	0.68	0.76	28716
weighted avg	1.0	1.0	1.00	28716

Table 4.2: Event Detection performance for Arabic documents in test set.

We note that the performance of event detection is better in English than Arabic. However, it is to be noted that for Arabic, the event detection task is extremely imbalanced (the ratio of positive documents is 0.2% vs 14.8% for English).

4.10.2 How good is the uncertainty/confidence characterization of the event detection model?

As mentioned above, the confidence score of the event detection is taken to be the predicted probability of the max class. Figure. 4.3 gives the precision vs confidence plot for English documents and Figure. 4.4 for Arabic documents.

From Figure. 4.3, we can see that Precision, Recall, F-1 metrics increase almost uniformly as the confidence threshold is increased. The Recall and F-1 metrics fall sharply for high values of the threshold. However for Arabic (From Figure. 4.4), there isn't much performance

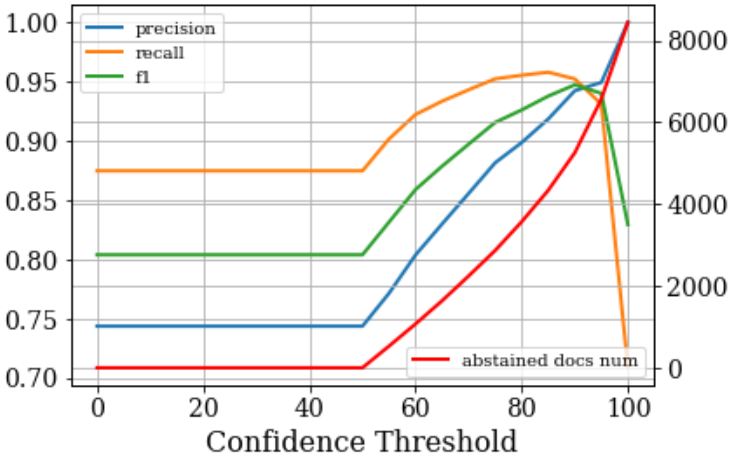


Figure 4.3: Precision vs Confidence plot for English. The right-side axis denotes the number of documents that will be abstained at the current threshold

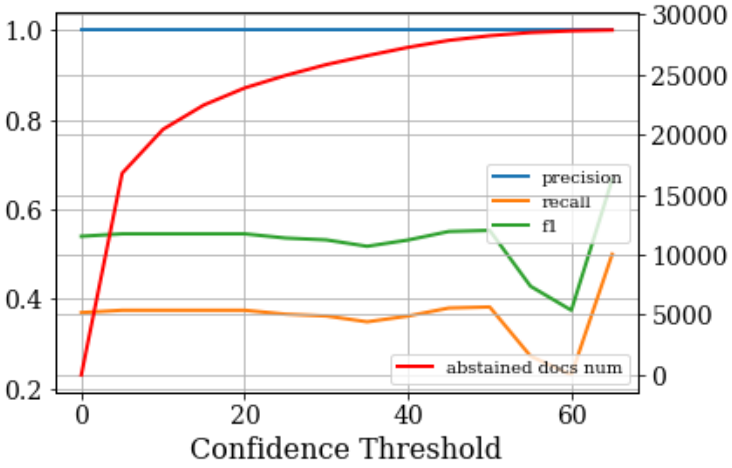
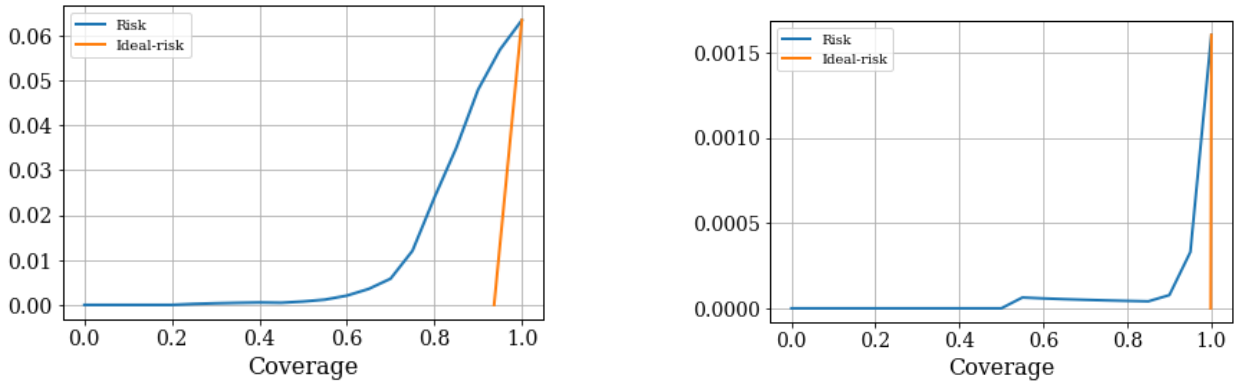


Figure 4.4: Precision vs confidence plot for Arabic



(a) (b)

Figure 4.5: Risk-coverage curve for (a) English and (b) Arabic. The orange curve represents the ideal-risk i.e., the risk when the classifier is correctly calibrated ($\text{loss}(x_i) > \text{loss}(x_j)$, iff $\text{conf}(x_j) < \text{conf}(x_i)$)

improvement achieved by thresholding on the confidence.

Another way at quantifying uncertainty of the classification model would be to look at the area under the Risk-Coverage (RC) curve. Figure. 4.5a provides the RC curve for English and Arabic.

We notice from Figure. 4.5a that the risk drops to almost zero at about 50% coverage. Note Risk is defined as the percentage of False classifications.

4.10.3 What is the performance of human annotators for each micro-task?

For estimating the performance of each human annotator, we randomly chose a micro-task for a given document and asked the annotator to provide encoding for the field denoted by the micro task. For each micro-task, the annotators were shown the encoding of the remaining fields, however they were not allowed to change these fields.

The pie chart shown in Figure. 4.6 gives the time distribution (in seconds) per micro-task.

The time per task reported here is the average of the median time taken by each annotator. The median is taken after removing the top and bottom 10 percentile of times of an annotator. This is done to remove anomalous records (For example, cases where the annotator simply leaves the autogsr web-page on without logging off.)

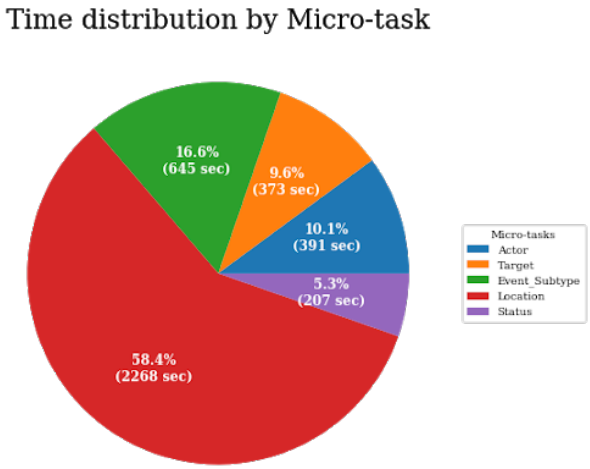


Figure 4.6: Average time spent per document per microtask by an annotator

Figure. 4.6 shows the human annotators spent maximum time (58.4%, approx 37 minutes) in encoding the location of the events. The least amount of time was spent on identifying the status of the article. Note status here refers to the event detection task. The average accuracy per micro-task across all annotators is provided in Figure. 4.7. We can see that the annotators have difficulty in identifying the status of an article.

4.10.4 How many documents are abstained per micro-task by the ML system?

In this section we discuss the number of documents the ML system abstained from making a prediction for each micro-task. The threshold for abstention for different micro-tasks was

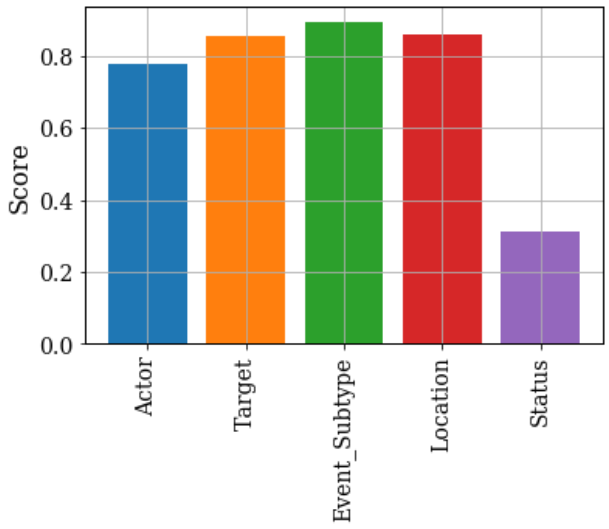


Figure 4.7: Average accuracy of human annotators on micro-tasks

Micro-Task	**#Abstained Documents**
Event Detection	1936 / 57401 (3.00%)
Location	307/2137 (14.3%)
Actor/Target	1002/2137 (46.8%)
All (Location, Actor, Subtype)	320/2137 (14.9%)

Table 4.3: Documents sent for human supervision

identified based on the validation set performance. The threshold for event detection for Table. 4.3 summarizes the number of documents we sent for human supervision for each micro-task

4.10.5 What is the event coding performance of the ML only System?

The table. 4.4 provides the quality scores of the events extracted from the machine learning model in the un-abstained set (i.e from documents that are not sent for human supervision). Note we score events extracted from a document only against the ground-truth events from

Metrics	Macro-avg	Micro-avg
#docs	888	888
# GroundTruthEvents	2039	2039
# ExtractedEvents	1981	1981
Actor Score	0.373888	0.602880
Date Score	0.940764	0.840740
Event Subtype Score	0.449895	0.680532
Location Score	0.866620	0.800837
Target Target Score	0.662297	0.677204
Target Status Score	0.590778	0.828618
Precision	0.691522	0.949447
Recall	0.720177	0.891757
Quality Score	2.774264	2.999131

Table 4.4: Performance metrics of ML only system

that document. For overall precision and recall we average precision and recall for each individual document.

We can see from Table. 4.4 that the ML only system is very good at location and date extraction. The system does not perform well on Actor identification. On further investigation of the events extracted by the ML system we found that the ML system is not able to distinguish between the different state actors of a given country. For example, ML model is not able to differentiate between “Iraqi Special Forces” and “Iraqi Security Forces” or “Iraqi Intelligence Service”. We thus performed another evaluation after combining all state actors belonging to a country into a single entity. That is, we group all state actors of a country like “Iraq Security Forces”, “Iraqi Intelligence Service” , “Iraqi Police”, “Baghdad International Airport Security”, “Iraqi Military” and “Iraqi Special Forces” into one single entity “Iraqi State Actors”. The performance of our system after this correction is shown in Table. 4.5.

Metric	Macro-avg	Micro-avg
#docs	888	888
# GroundTruthEvents	2039	2039
# ExtractedEvents	1981	1981
Actor Score	0.551631	0.727570
Date Score	0.940403	0.842904
Event Subtype Score	0.451160	0.682170
Location Score	0.867125	0.805017
Target Target Score	0.588879	0.677740
Target Status Score	0.661664	0.822260
Precision	0.691522	0.951026
Recall	0.720177	0.893247
Quality Score	2.892902	3.087752

Table 4.5: Performance metrics of ML only system with partial entity matching. Here different state actors of one country grouped into one entity

Metric	Macro-avg	Micro-avg
#docs	2137	2137
# GroundTruthEvents	4265	4265
# ExtractedEvents	4239	4239
Actor Score	0.621032	0.755090
Date Score	0.952899	0.878390
Event Subtype Score	0.539359	0.646776
Location Score	0.926107	0.838040
Target Target Score	0.680881	0.703337
Target Status Score	0.727046	0.828054
Precision	0.791948	0.946720
Recall	0.819668	0.915887
Quality Score	3.121909	3.161470

Table 4.6: Performance metrics of the hybrid system for MANSA Event Encoding

4.10.6 What is the event coding performance of the hybrid system?

Table. 4.6 provides the event extraction performance for the Hybrid system. The hybrid system includes predictions from the ML system along with inputs from human annotators for the micro-tasks on documents where the ML system was not confident on.

4.10.7 What is the overall Conclusion for creating a MANSA event encoding system?

From Tables. 4.4 and 4.5 we can see that the ML models performs well in identifying Location, Date, and event status as compared to the human annotators. The human annotators on average perform better than the ML model on Actor and Target identification. A viable system could be one where the ML system performs the location, date and status identification and human annotators provide the actor/target extraction. The human annotators on average take approximately 12.6 minutes per document for actor and target identification and have an accuracy of approx 80% for the extraction of the two fields. So for a dataset similar to our test set, we would need approximately 448.77 man hours.

4.11 Discussion

We have presented an event encoding system for Military Action and Non-state actor events as well as for Civil Unrest events. Specifically we built a system to minimize human effort and to judiciously decide suitable human-machine combinations to yield high performance. The innovative aspects of our system are as follows:

- Hybrid Event encoding wherein the machine learning system decides for which documents and which fields human supervision is required.
- State-of-art Geocoding, wherein location entities in an article are disambiguated based on a language independent supervised classification engine.
- Unsupervised extraction of semantic information of type <*actor, target, type, location, date*> from text using universal dependency parsing and multilingual wordnet.

.

Chapter 5

Conclusion and Future Work

In this dissertation we have summarized our experiences from operating an open source forecasting system called EMBERS [45]. We introduced techniques for online drift adaption once a change point is detected. Based on the lessons learnt, we understand that an automated event coding system is essential to ensure proper functioning and success of event forecasting. However, based on previous works [30] we find that fully automated event coding systems are not quite reliable and hence we have outlined a system that tries to automate event coding whenever possible and when a decision cannot be made with certainty learns to send the document/article for human supervision. Then we presented an approach for event detection with the capability to say 'I don't know' when it is uncertain. We showcase how the sentence level uncertainty quantification can offer insights about the reasons for uncertain predictions by a model. Next we introduce a state-of-the-art geocoding pipeline for identifying and grounding (attaching to corresponding latitude and longitude) locations mentioned in text. Finally we build a automated event extraction framework with uncertainty quantification at all levels. Having uncertainty quantification at all levels of the proposed framework enables to efficiently direct human effort (via. human intervention) only on uncertain micro-tasks,

thereby minimizing the overall human effort.

5.1 Ethics Considerations

In this dissertation we elucidate about anticipatory intelligence systems and automated approaches to ground-truth creation necessary for such systems. Such systems have many powerful legitimate uses but are also susceptible to abuse. Events like civil unrest enhances the ability of citizens to communicate not only their views but also their priorities to those who govern them. An open sources indicators approach, as we have used here, is a potentially powerful tool for understanding the social construction of meaning and its translation to behaviour. An anticipatory intelligence system and an ground-truth event extraction system can contribute to making the transmission of citizen preferences to government less costly to the economy and society as well. There are economic costs to even peaceful disruptions embodied in events like civil unrest due to lost work hours and the deployment of police to manage traffic and the interactions between protesters and bystanders.

The potential power of an anticipatory intelligence System, like those of most scientific advances, is susceptible to abuse by governments. The appropriate safeguards require developing transparent and accountable democratic systems, not outlawing science.

Similarly, automated (or hybrid) ground-truth extraction systems such as the event extraction framework presented in this thesis allow one to study and understand the occurrence of an event in relation to other events in the past and uncover potential patterns or causes. Such systems provide the ability to scale-up to the ever increasing amount of information being churned out from the web but at the same time reducing cost significantly when compared to manual or partially automated approaches. Though in this thesis we make an effort to ensure the model is well calibrated and that the model learns to know what it doesn't know,

being a machine learning approach, it still may make incorrect predictions or miss certain types of events. Finally the quality of extractions from such systems are also based on the validity and quality of data used for training. Thus it is possible that such systems can be biased towards certain kinds of actors or targets. Therefore, it is essential for the end user to understand that the extractions of such systems are still predictions from a machine learning model which must not be used directly in decision making but rather as an additional source of evidence.

5.2 Future Work

Future research directions are discussed below.

- **Multi-sensor belief fusion for understanding record level uncertainty.** Currently each individual attribute of an event record is extracted using a different pipeline (micro-tasks) independent of the others. Historical records can provide prior belief for different parts of the record occurring together. There is possibility that we can get highly confident individual field extractions but the combination of two or more extracted field values may be less likely. Thus moving away from just using singleton beliefs and trying to understand overall record level belief will improve the performance of event extraction systems.

- **Question answering for generalized event extraction.** Our current approach to event extraction though more scalable and easy to train, still requires significant effort to get started with a new type of event. Ground-truth data needs to be generated manually to create the initial training data. This can be expensive and time consuming. Recent advances in Question Answering systems show promise in ability to generalize

to generic queries and thus QA systems can be used to identify new types of events without any training data.

Bibliography

[1] A. Alexandari, A. Shrikumar, and A. Kundaje. Selective classification via curve optimization. *arXiv preprint arXiv:1802.07024*, 2018.

[2] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840, 2008.

[3] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013.

[4] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, 2014.

[5] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward. Icews coded event data. *Harvard Dataverse*, 12, 2015.

[6] J. Cadena, G. Korkmaz, C. J. Kuhlman, et al. Forecasting Social Unrest using Activity Cascades. *PLOS One*, 10(6):e0128879, 2015.

[7] C. Center. The social, political and economic event database project (speed), 2016.

[8] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, et al. Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions. In *SIAM International Conference on Data Mining, April 24-26, 2014*, pages 262–270, 2014.

[9] P. Chakraborty, S. Muthiah, R. Tandon, and N. Ramakrishnan. Hierarchical quickest change detection via surrogates. *arXiv preprint arXiv:1603.09739*, 2016.

[10] A. P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

[11] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[12] R. Dingledine, N. Mathewson, and P. F. Syverson. Tor: The Second-Generation Onion Router. In *USENIX Security Symposium, August 9-13, 2004*, pages 303–320, 2004.

[13] L. Dong and M. Lapata. Coarse-to-fine decoding for neural semantic parsing. *arXiv preprint arXiv:1805.04793*, 2018.

[14] A. Doyle, G. Katz, K. Summers, C. Ackermann, et al. Forecasting Significant Societal Events using The EMBERS Streaming Predictive Analytics System. *Big Data*, 2(4):185–195, dec 2014.

[15] G. F. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio. Large margin deep networks for classification. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 850–860, USA, 2018. Curran Associates Inc.

[16] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[17] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in neural information processing systems*, pages 4878–4887, 2017.

[18] Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pages 2151–2159, 2019.

[19] J. A. Goldstone, R. H. Bates, D. L. Epstein, et al. A Global Model for Forecasting Political Instability. *American Journal of Political Science*, 54(1):190–208, 2010.

[20] J. A. Goldstone, R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward. A global model for forecasting political instability. *American Journal of Political Science*, 54(1):190–208, 2010.

[21] J. Gordon and E. H. Shortliffe. The dempster-shafer theory of evidence. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, 3:832–838, 1984.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

[23] L. Hamers et al. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3):315–18, 1989.

[24] A. Hoegh, S. Leman, P. Saraf, and N. Ramakrishnan. Bayesian Model Fusion for Forecasting Civil Unrest. *Technometrics*, 57(3):332–340, February 2015.

[25] F. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong. An overview of event extraction from text. In *conference; ISWC 2011; 2011-10-23; 2011-10-23*, pages 48–57. CEUR-WS. org, 2011.

[26] H. Jiang, B. Kim, M. Guan, and M. Gupta. To trust or not to trust a classifier. In *Advances in neural information processing systems*, pages 5541–5552, 2018.

[27] A. Jøsang. *Subjective logic*. Springer, 2016.

[28] A. Kamath and R. Das. A survey on semantic parsing. *arXiv preprint arXiv:1812.00978*, 2018.

[29] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[30] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan. Detecting and Forecasting Domestic Political Crises: A Graph-Based Approach. In *ACM Web Science Conference, WebSci '14, June 23-26, 2014*, pages 192–196, 2014.

[31] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan. Combining Heterogeneous Data Sources for Civil Unrest Forecasting. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM, August 25 - 28, 2015*, pages 258–265, 2015.

[32] K. Leetaru and P. Schrodt. GDELT: Global Data on Events, Location and Tone, 1979–2012. *ISA Annual Convention*, pages 1979–2012, 2013.

[33] K. Leetaru and P. A. Schrodt. Gdelt: Global data on events, location, and tone. In *ISA Annual Convention*. Citeseer, 2013.

[34] L. Li, M. L. Littman, T. J. Walsh, and A. L. Strehl. Knows what it knows: a framework for self-aware learning. *Machine learning*, 82(3):399–443, 2011.

[35] Z. Liu, Z. Wang, P. P. Liang, R. R. Salakhutdinov, L.-P. Morency, and M. Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems*, pages 10622–10632, 2019.

[36] H. Llorens, L. Derczynski, et al. TIMEN: An Open Temporal Expression Normalisation Resource. In *International Conference on Language Resources and Evaluation, LREC, May 23-25, 2012*, pages 3044–3051, 2012.

[37] A. Mahendiran, W. Wang, J. A. S. Lira, et al. Discovering Evolving Political Vocabulary in Social Media. In *International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC, October 30 - November 1, 2014*, pages 26–32, 2014.

[38] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[39] S. Muthiah, B. Huang, J. Arredondo, et al. Planned Protest Modeling in News and Social Media. In *AAAI Conference on Artificial Intelligence, January 25-30, 2015*, pages 3920–3927, 2015.

[40] T. H. Nguyen, K. Cho, and R. Grishman. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, 2016.

[41] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM, 2005.

[42] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM, 2005.

[43] S. P. O'Brien. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12(1):87–104, 2010.

[44] J. PLATT. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classiers*, 1999.

[45] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, et al. 'Beating the news' with EMBERS: Forecasting Civil Unrest using Open Source Indicators. In *International Conference on Knowledge Discovery and Data Mining, KDD, August 24 - 27, 2014*, pages 1799–1808, 2014.

[46] T. Rekatsinas, S. Ghosh, S. R. Mekaru, et al. SourceSeer: Forecasting Rare Disease Outbreaks using Multiple Data Sources. In *SIAM International Conference on Data Mining, April 30 - May 2, 2015*, pages 379–387, 2015.

[47] P. Saraf and N. Ramakrishnan. Embers autogsr: Automated coding of civil unrest events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 599–608. ACM, 2016.

[48] P. A. Schrodt. Tabari: Textual analysis by augmented replacement instructions. *Dept. of Political Science, University of Kansas, Blake Hall, Version 0.7. 3B3*, pages 1–137, 2009.

[49] P. A. Schrodt, J. Beieler, and M. Idris. Three'sa charm?: Open event data coding with el: Diablo, petrarch, and the open event data alliance. In *ISA Annual Convention*. Citeseer, 2014.

[50] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

[51] J. Strötgen and M. Gertz. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, 2010.

[52] N. N. Taleb. *The Black Swan. The Impact of the Highly Improbable.* Random House Inc., 2008.

[53] S. Thulasidasan, T. Bhattacharya, J. Bilmes, G. Chennupati, and J. Mohd-Yusof. Combating label noise in deep learning using abstention. In *International Conference on Machine Learning*, pages 6234–6243, 2019.

[54] W. Wang, R. Kennedy, D. Lazer, and N. Ramakrishnan. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503, 2016.

[55] M. D. Ward, N. W. Metternich, C. Carrington, et al. *Geographical Models of Crises: Evidence from ICEWS*, pages 429–438. CRC Press, Boca Raton, FL, 2012.

[56] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[57] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[58] X. Zhang, F. Chen, C.-T. Lu, and N. Ramakrishnan. Mitigating uncertainty in document classification. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, June 2019.

[59] X. Zhang, F. Chen, C.-T. Lu, and N. Ramakrishnan. Mitigating uncertainty in document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3126–3136, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[60] L. Zhao, F. Chen, et al. Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling. *PLOS ONE*, 9(10):e110206, 2014.

[61] L. Ziyin, Z. Wang, P. P. Liang, R. Salakhutdinov, L.-P. Morency, and M. Ueda. Deep gamblers: Learning to abstain with portfolio theory. *arXiv preprint arXiv:1907.00208*, 2019.