Event Categorization and Key Prospect Identification from Storylines

Manu Shukla[1,2,3](✉), Andrew Fong[1,2,3], Raimundo Dos Santos[1,2,3], and Chang-Tien Lu[1,2,3]

[1] Omniscience Corporation, Palo Alto, CA, USA
{manu.shukla,andrew.fong}@omni.sc
[2] US Army Corps of Engineers ERDC GRL, Alexandria, VA, USA
raimundo.f.dossantos@erdc.dren.mil
[3] Computer Science Department, Virginia Tech, Falls Church, VA, USA
ctlu@vt.edu
http://www.omni.sc

Abstract. Event analysis and prospect identification in social media is challenging due to endless amount of information generated daily. While current research focuses on detecting events, there is no clear guidance on how those events should be processed such that they are meaningful to a human analyst. There are no clear ways to detect prospects from social media either. In this paper, we present DISTL, an event processing and prospect identifying platform. It accepts as input a set of storylines (a sequence of entities and their relationships) and processes them as follows: (1) uses different algorithms (LDA, SVM, information gain, rule sets) to identify themes from storylines; (2) identifies top locations and times in storylines and combines with themes to generate events that are meaningful in a specific scenario for categorizing storylines; and (3) extracts top prospects as people and organizations from data elements contained in storylines. The output comprises sets of events in different categories and storylines under them along with top prospects identified. DISTL uses in-memory distributed processing that scales to high data volumes and categorizes generated storylines in near real-time.

1 Introduction

In social media channels like Twitter, emerging events propagate at a much faster pace than in traditional news. Combining relevant facts together (while discarding unimportant ones) can be very challenging because the amount of available data is often much larger than the amount of processing power. This implies that many systems are unable to keep up with increasingly large volumes of data, which may cause important information to be missed. Event processing, therefore, is at a minimum dependent on two tasks: (1) collecting all the facts, entities, and their relationships; (2) grouping them by their themes of discussion, space, and timeframes. These two tasks should be performed in a distributed paradigm for maximum coverage. In the real world, not every piece of information

C. Grueau et al. (Eds.): GISTAM 2016, CCIS 741, pp. 62–88, 2017.
DOI: 10.1007/978-3-319-62618-5_5

can be thoroughly investigated in a timely manner. The goal, therefore, is to maximize the two previous tasks so that an event can be described with the most number of pertinent facts that yields the most complete picture. Figure 1 provides a visual representation of the idea. The figure shows seven tweets with a connection to the Boston area: t2, t3, and t7 are related to the Boston Marathon Bombings of April 2013, while t1 and t5 are about baseball, and t4 and t6 are about finance. First, these messages are certain to come hidden among millions of other tweets of different natures. Further, they relate to different topics, which indicates they should be presented separately. As seen in Fig. 1, all of the tweets are first transformed into simple storylines, and then grouped into three different themes ("Boston Marathon Bombings", "Wall Street News" and "Boston Red Sox"), which may be better suited to present to different audiences.

Social media is also a key place to identify emerging prospects for investment, partnerships and acquisitions [31]. For key prospect identification, finding top people and organizations from storylines model or performing unsupervised learning can be crucial for analyst. Figure 2 provides a visual representation of the idea. The figure shows three tweets (t1, t2, and t3) related to connected cars, self driving cars and industry activity. As Fig. 2 shows, all of the tweets are first transformed into simple storylines, and then four different prospects identified, three prospects represent large companies (Alibaba, BMW and Intel) and one (Mobileye) is a small company. The latter may be of particular interest to an analyst looking to invest in the connected car business, while other three may be better suited for analyst tracking large investments in the area.

The goal of this paper is to perform the above tasks using DISTL, Distributed In-memory Spatio-Temporal storyLine categorization platform (also shown in Fig. 1) [29], a system that ingests storylines derived from tweets, and allocates them to appropriate events along with finding top prospects for investments and acquisition in a given domain. The criteria used for the allocation process is that storylines have common themes, are located in nearby areas, and take place during close timeframes. DISTL uses as input the storylines generated by DISCRN [22], and is an in-memory spatio-temporal event processing platform that can scale to massive amounts of storylines using Big Data techniques. The platform helps analysts find faint, yet crucial events by separating storylines into groups, which allow analysts to sift through them in subsets under specific contexts.

A storyline is simply a time-ordered connection of facts that take place in a geographical area. In Fig. 1, for example, "police → block off → downtown Boston" represents a simple storyline related to a bigger event (the Boston Marathon Bombings). Storylines may be variable in length, and made as elastic as desired. In this paper, we do not show how these storylines are generated. Rather, we refer the reader to our previous work, DISCRN, which is a distributed platform specifically dedicated to generating storylines.

In order for a storyline "to be told", the user must first select a starting entity, such as a person or organization, from where the story can investigated. By checking the connections from that starting entity to other entities, one can then combine the facts together into a bigger event. For example, one may select a "person carrying a back pack" from one tweet to be the starting entity, and obtain other facts from other tweets, such as "entering subway", and "making a phone call", which would paint a more complete picture of a possible crime. DISCRN is a distributed system that mines storylines, as described above, at scale. It is effective in extracting storylines from both short unstructured tweet snippets and structured events such as in GDELT (Global Databases of Events, Language and Tone) [13]. DISCRN uses MapReduce [9] to generate all storylines from a specified starting entity from a large set of tweets. Since MapReduce is disk-based, it becomes less than ideal for highly-iterative event processing algorithms used in DISTL. For that reason, it is imperative to explore memory-based solutions explained later.

The key contributions of the platform are:

- **Design a Framework to Generate Spatio-temporal Events and Investment Prospects from Storylines.** Multiple algorithms (LDA, Information gain, classification) are applied to determine events that incorporate themes, location and time elements. The framework also identifies prospects for investment, partnership or acquisition for a domain.
- **Distribute In-memory Event Processing and Prospect Identification.** In-memory distributed architecture scales highly iterative event and prospect generation. This allows for processing large number of storylines efficiently.
- **Develop Rules based Technique to Categorize Storylines into Events.** Rules allow user fine-grained control on incorporating storylines into events. This provides user flexibility in focusing on theme, location or time.
- **Conduct Extensive Experiments to Validate Events Categorization and Top Prospects Identification from Storylines.** Framework and algorithms are validated with extensive experiments in multiple domains. The results effectively categorized storylines under meaningful events and found useful prospects for investigation.

The rest of the paper is organized as follows. Section 1 provides an overview of storylines and event creation with them. Section 2 describes the related works on event creation from social media data. Section 3 describes the techniques in DISTL in detail and Sect. 4 describes the architecture of components used to perform it. Section 5 presents experiments performed with datasets on different storyline subjects and themes on which meaningful and interesting events and top prospects were generated. Section 6 provides conclusions of the study.

2 Related Works

This section provides related works in event creation and distributed in-memory algorithms. Event creation in social media is a widely researched field. Event creation consists of identifying the event and characterizing it. Previous work primarily focuses on detecting events instead of categorizing elements under them.

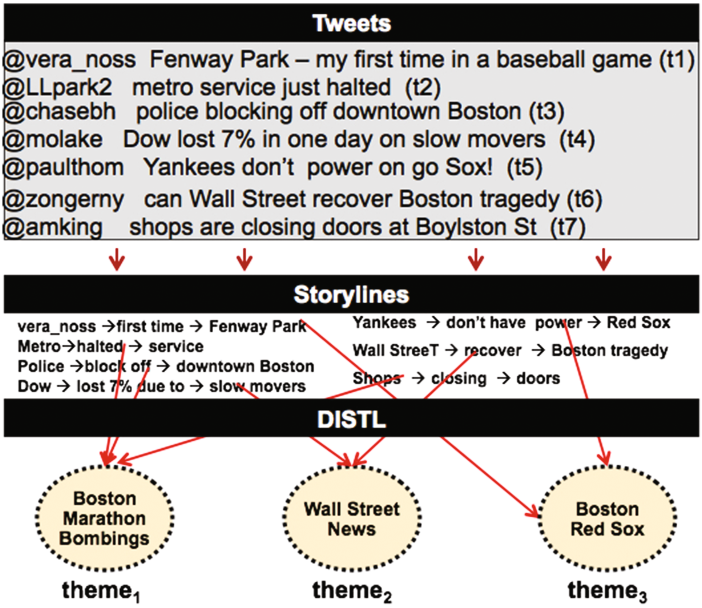
Tweets
@vera_noss Fenway Park – my first time in a baseball game (t1)
@LLpark2 metro service just halted (t2)
@chasebh police blocking off downtown Boston (t3)
@molake Dow lost 7% in one day on slow movers (t4)
@paulthom Yankees don't power on go Sox! (t5)
@zongerny can Wall Street recover Boston tragedy (t6)
@amking shops are closing doors at Boylston St (t7)
Storylines
vera_noss →first time → Fenway Park
Metro→halted → service
Police →block off → downtown Boston
Dow → lost 7% due to → slow movers
Yankees → don't have power → Red Sox
Wall StreeT → recover → Boston tragedy
Shops → closing → doors
DISTL
Boston Marathon Bombings
theme1
Wall Street News
theme2
Boston Red Sox
theme3

Fig. 1. Events used to categorize storylines.

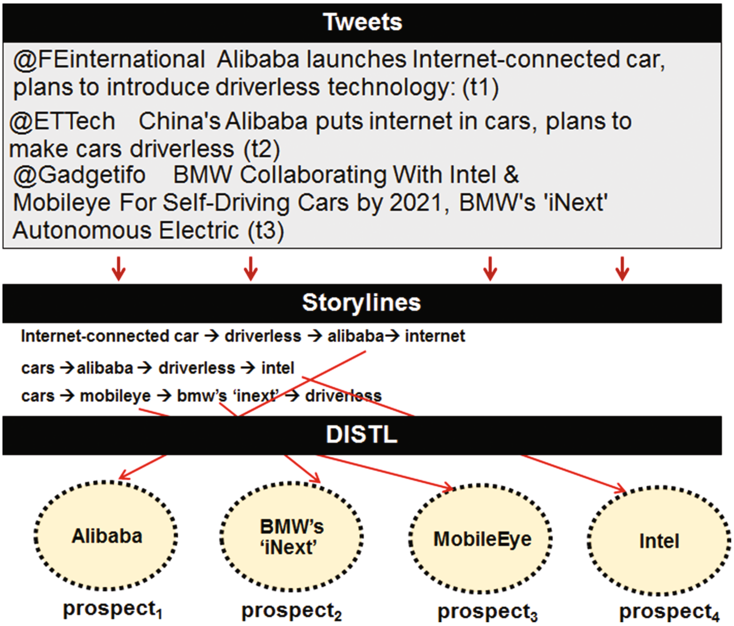
Tweets
@FEinternational Alibaba launches Internet-connected car, plans to introduce driverless technology: (t1)
@ETTech China's Alibaba puts internet in cars, plans to make cars driverless (t2)
@Gadgetifo BMW Collaborating With Intel & Mobileye For Self-Driving Cars by 2021, BMW's 'iNext' Autonomous Electric (t3)
Storylines
Internet-connected car → driverless → alibaba→ internet
cars →alibaba→ driverless → intel
cars → mobileye → bmw's 'inext' → driverless
DISTL
Alibaba
prospect1
BMW's 'iNext'
prospect2
MobileEye
prospect3
Intel
prospect4

Fig. 2. Prospects from tweets used to build storylines of interest.

A useful survey of detecting events in social media has been preformed [11]. Inferential models are explored for detecting events from news articles [18]. Graphical models are used to detect events from multiple user posts over social media streams [27]. Non-parametric based scan statistic based event creation as anomalous subgraphs is performed [7]. Dynamic query expansion and anomaly creation are combined to detect events [25]. Clustering techniques along with signature of known events based supervised event creation is proposed [1]. Wavelets on frequency based raw signals and modularity based graph partitioning on transformed space has been shown to be effective [10]. Clustering based sub-event identification is investigated [17]. Tweets as replacement of news is explored [16]. Segment based event detection in tweets is proposed [14].

Spatio-temporal event creation has also attracted attention as events tend to be localized phenomenon. Jointly modeling structural contexts and spatio-temporal burstiness is used in event forecasting [26]. Burstiness and spatio-temporal combinatorial patterns for localizing events spatially is explored [12]. Classifier based event creation along with spatio-temporal model to determine event trajectory [21]. A visual analytics way to detect spatio-temporal events using LDA based topics and outlier creation is explored [6]. A sequential spatio-temporal event detection system from tweets is proposed [15].

NoSQL database based event detection techniques using clustering is explored [23]. Scalability in event creation is usually achieved through transforming the problem to efficient domains. Scalability in event creation from social media streams with event based clustering by reducing problem to a record linkage problem is investigated [19]. A scalable non-negative matrix factorization based technique to detect events in social media is presented [20]. However, in case of storytelling, events have to be generated such that all storylines are attributed under the event making it imperative that none are dropped. That requires scaling through distribution rather than problem transformation. Machine learning has been used widely to manage investments. Financial Genetic Programming (FGP) is used for investment decision making [28]. Neural nets have been used in predicting takeover targets [30].

There are no known techniques for distributed event creation. DISTL applies highly iterative techniques to event theme generation that can not be scaled efficiently with disk based distribution such as MapReduce. Use of Apache Spark to perform topic modeling, entity selection and classification in memory allows for much more efficient scaling. It distributes the entire sequence of steps starting from composite event generation and subsequent storyline categorization into those events in-memory. This allows to scale the process completely and maximize impact of distribution.

3 Event Generation and Prospect Identification Techniques

In this section the techniques used to generate events from storylines and categorize storylines under those events are described. Subsection 3.1 provides brief overview of distribution techniques in Spark. Subsection 3.2 presents theme generation technique followed by Subsect. 3.3 that explains how events are generated from themes and storylines assigned to the events. Subsection 3.4 describes the top prospect identification techniques in DISTL.

3.1 In-Memory Distribution in Spark

Apache Spark is an in-memory distribution framework that allows computations to be distributed over a large number of nodes in a cluster [24]. The programming constructs available in Spark represent transformation of data on disk into RDDs (Resilient Distributed Datasets), which reside in-memory. Operations applied on the RDDs to generate values that can be returned to the application. RDDs provide fault tolerance in case one or more nodes of the cluster fail. The algorithms typically useful for Spark are the ML (Machine Learning) and statistical functions that are highly iterative in nature. Performing highly distributed operation in any other distributed construct such as MapReduce is computationally expensive due to data written to disk in each iteration. Spark allows for highly efficient iterative operations as they are all performed in memory.

The main operations provided by Spark that allows it process large volumes of data in parallel can be broadly categorized into *actions* and *transforms* [4]. The *transform* operations commonly used include *map*, *filter*, *flatMap*, *join*, *reduceByKey* and *sort*. The *action* operations commonly used are *reduce*, *collect* and *count*. The *map* operation returns a distributed dataset by applying a user specified function to each element of a distributed dataset. A *flatMap* does the same except one input term can be mapped to multiple output items. *reduceByKey* operation aggregates the values of each key in key-value pairs <K,V> according to provided reduce function. *filter* returns datasets from source for which given function evaluates true. Spark also allows to keep a read-only value of variables in cache on each node instead of shipping them with each task through broadcast variables.

3.2 Theme Creation

Several major theme recognition techniques are made available to the analyst. The event creation technique uses top-weighted keywords as themes. A dictionary-based method assigns storylines to event buckets. Rule based storylines categorization is performed. We can generate events based on theme, location and time. The dictionary is generated for themes by analyzing the terms of the storylines and discovering key ones. The recognized themes are then used to categorize the storylines.

The key aspect of event generation is identifying the entities that are closest to significant events. The sequence of steps in events generation and assigning storylines to events is shown in Fig. 3. The flow consists of 3 main steps; process storylines, build themes and create events, and score and categorize storylines. The first step processes storylines and identifies spatial and temporal entities in them. Here supervised and unsupervised techniques are used in the identification of the most critical entities, which are used in the subsequent step. Theme, spatial and temporal entities are combined to generate events. The storylines are then categorized under the events in the last step of flow. The three algorithms used in second step of flow are as follows.

- Topic Modeling based event creation: Topic models are generated using Latent Dirichlet Allocation (LDA). It uses a list of storylines as input data, where each storyline represents a document. This technique outputs a number of topics specified by the user, where each topic is comprised as a set of keywords.
- Feature Selection through Information Gain based event creation: This technique extracts top n keywords by information gain from storylines. Each storyline is treated as a document. Each of the highest n information gain value keywords is treated as belonging to the subject for which labeled data was generated.
- Classifier based event creation: This technique uses a classifier trained with user generated set of storylines for a particular subject. This model is then used to classify storylines into ones belonging to that subject or not. An example would be if analyst wants to separate all storylines that are related to earnings of a company from ones that are not. A classifier based technique works best in case of known subjects being analyzed in storylines. Events under which storylines are categorized are generated using most frequent theme, location and time entities in positively labeled training data. Storylines are also scored based on the classifier score or the score of their entities for topic or feature selection. These scores are then used to filter top storylines by applying a threshold. Data elements for entities of top storyline are then identified and used to create top people and organizations as prospects.

Topic modeling falls under unsupervised learning while other two (information gain and classifier) are supervised. They require training data in order to generate the themes. All these techniques are highly iterative and under large datasets computationally expensive especially in terms of building model. Algorithm 1 shows the application of 3 techniques to categorize events. Step 1 performs the extraction of entities from storylines and generating RDDs of storylines from JSON output produced by DISCRN. One RDD is created for training data and one from scoring data. A combined index of entities is generated. Step 2 then generates RDDs of theme entities and other entities identified as location and time as PairRDDs. PairRDDs are <Key, Value> pair representation in RDDs. It then performs LDA based topic modeling, feature selection based on information gain or SVM model generation and most frequent entities from positively labeled training data to extract themes from entities. All operations are implemented such that they are performed in memory.

3.3 Event Generation and Storylines Assignment

Events are generated by combining themes with the spatial and temporal entities identified in storylines. Algorithm 2 shows how generating events based on

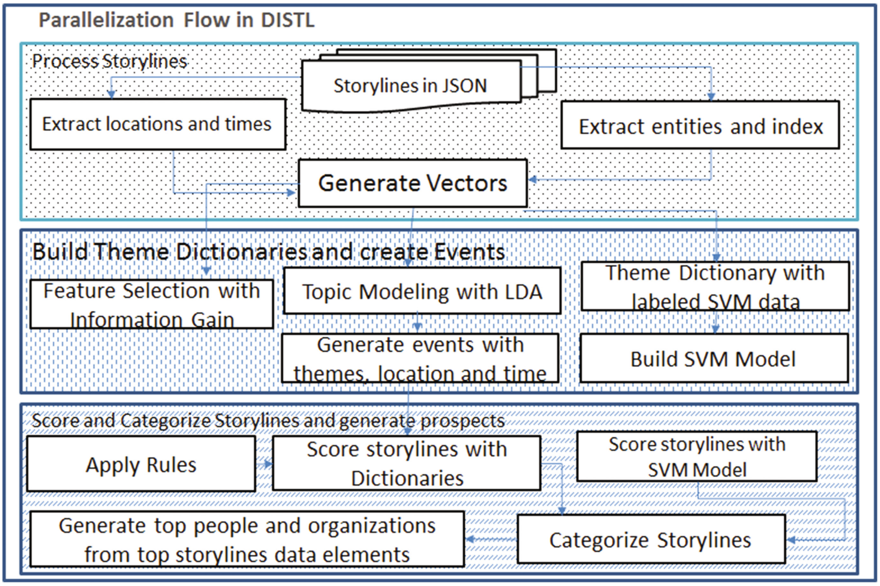
Parallelization Flow in DISTL
Process Storylines
Storylines in JSON
Extract locations and times
Extract entities and index
Generate Vectors
Build Theme Dictionaries and create Events
Feature Selection with Information Gain
Topic Modeling with LDA
Theme Dictionary with labeled SVM data
Generate events with themes, location and time
Build SVM Model
Score and Categorize Storylines and generate prospects
Apply Rules
Score storylines with Dictionaries
Score storylines with SVM Model
Generate top people and organizations from top storylines data elements
Categorize Storylines

Fig. 3. DISTL system flow.

themes, location and time entities is performed in-memory with the location, time and theme entities extracted from entities RDD and then combined together to create events. The task of finding the combinations of location, time and entities based on one or more subject depends on the technique used for subject creation or labeled data based entity extraction. These are crucial to identifying events and associating entities with events. Step 2 categorizes storylines into events. This approach tests keywords in a storyline against the spatial, temporal and theme entities and assigns it to the theme based events using rules specified by user.

The rules provided by user to categorize storylines into events with theme, location and time elements and their application on storylines is explained below.

Categorization Rules Format. The rules are of following format:
theme ([$and|or$] location [$and|or$] time)
Hence the rules can take any of the following forms:

1. theme *or* (location *and* time)
2. theme *or* location
3. location
4. time

The rules specify which entities in a storyline need to match with rule entity of particular type in order to associate a storyline to the event. Hence Rule 1 specifies that only if storyline entities match either theme or location and time

Algorithm 1. Generate themes.

```
Input: $\{storyline_i\}$,$\{labeled_j\}$ {unlabeled storylines and labeled storylines for supervised learning}
Output: $\{event_k$,$storyline_i$ under each $event_k\}$
    {each event definition and storylines under each event}
{step 1: parse storylines and extract entities to generate labeled and unlabeled RDDs}
Create trainingDataRDD from labeled storylines file on distributed file system using map transform
Create entityIndexRDD as index of entities to integers using flatMap and filter transforms
Create testingDataRDD from unlabeled storylines file on distributed file system using map transform
Create labeledVectorsRDD and unlabeledVectorsRDD with vectors for storylines using zipwithindex and distinct transforms
{step2: Identify location and time entities }
Extract location and time entities from all entities and build locationTimeRDD using flatMapToPair transform
as PairRDD as map of entities and their type location or time
{step3: run LDA model and get topics or SVM Modeling or Info Gain Feature Selection}
if technique is topic modeling then
    ldaModel = new LDA().setK(noOfTopics).run( storylineEntityVectorsRDD);
else if technique is feature selection based on information gain then
    Create FeatureSelection featureSelection object
    Perform MRMR feature selection by featureSelection.minimumRedundancyMaximumRelevancy
    (storylineEntityVectorsRDD, numberOfFeatures);
    extract information gain values by featureSelection.relevancyRedundacyValues();
else if technique is classifier based then
    {call svm classification routine}
    Build SVMModel by invoking SVMWithSGD.train(labelPointsRDD, numIterations)
    score testingDataRDD with model
    Extract themes as PairRDD¡String,Integer¿ from positively scoring data
end if
```

Algorithm 2. Generate Events.

```
Input: $\{storyline_i\}$,$\{labeled_j\}$ {storylines and labeled storylines for supervised learning}
Output: $\{event_k\}$
    {each event definition and storylines under each event}
{step 1: Use themes and dictionaries generated in previous algorithm}
Get locTimeRDD from previous step
Get labeledVectorsRDD from previous step
{step 2: Use output from applied algorithm from previous step}
if technique used is topic modeling then
    {Applying top LDA weighted themes, locations and times}
    for all topic $\in$ Topics do
        Extract top location, time and theme term along with their weights
        Combine top weighted theme, time and location entity into event
    end for
    Get $k$ events where $k$ were number of topics extracted
else if technique is feature selection based on information gain then
    {Generate events with top info gain entities}
    Generate event as combination of top information gain theme, location and time
else if technique is classifier based then
    {Generate events with top positively labeled storylines location, time and theme entities by frequency}
    Calculate frequency of entities in positively labeled documents
    Combine top location, time and theme entities into events
end if
```

then categorize the storyline to the event. Rule 2 can categorize a storyline to the event if any of its entities matches either theme or location of the event while Rule 4 associates any storyline whose entities match the temporal entity of the event.

Rules Application. As each of these rules are applied to a storyline for each event, if any rule is satisfied for a storyline against an event, the storyline is categorized under that event. Algorithm 3 categorizes storylines under events applying the rules. Rules are broadcast to all the nodes and the storylines RDD then has each storyline in it run through the rules and associated with an event if any rule matches the storyline to the event. As soon as a storyline is associated with an event the rules application ends. Based on number of entities in a storyline that match rule's theme, location or time, a weight is assigned to storylines. For classifier events the weights are normalized with the storylines classifier score.

3.4 Prospect Generation from Storylines

Prospects are generated by extracting key people and organization from data elements used to build storylines. Algorithm 4 shows how generating prospects from tweets of storylines entities is performed in-memory with the tweets from entities RDD. The task of finding top scoring storylines utilizes the weights from topic modeling and features selection and the storylines scores against SVM classifiers as shown in step 2. These are then utilized to generate tweets from top storyline entities and extract top people and organizations as prospects in step 3.

4 System Architecture

This section describes the overall architecture of DISTL. The Subsect. 4.1 describes theme and dictionary creation component while Subsect. 4.2 describes the component that categorizes storylines into events. Prospect identification component is detailed in Subsect. 4.3.

4.1 Theme and Dictionary Creation

The system architecture of the platform to generate events in storylines is shown in Fig. 4. Due to large number of storylines generated from tweets collected on topics, the amount of data to be processed to generate events on the entities can be large. Event creation is performed as an extension to the DISCRN platform. In-memory distribution is essential to computing topics and perform feature selection based on information gain as these techniques tend to be highly iterative and do not scale well on disk based distribution paradigms such as MapReduce as disk I/O will be highly detrimental to performance. The modules described in this subsection generate themes from storylines, identify location and time entities and combine them to create composite events.

– Process Storylines: This job in Spark reads the storylines in parallel and extracts entities from them. Vectors are built with indices of entities in storylines.

Algorithm 3. Categorize storylines under Events.

```
Input: $\{storyline_i\}$,$\{event_j\}$,svmModel {storylines and labeled storylines for supervised learning}
Output: $\{event_k$,$storyline_i$ under each $event_k\}$
    {each event definition and storylines under each event}
{step 1: parse rules}
Broadcast rules to all worker nodes
Read rules in broadcast var
{step 2: Apply rules to generate events depending on algorithm previously applied}
if technique used is topic modeling then
    {Categorize storylines under topic events}
    for all topic $\in$ Topics do
        PairRDD<Integer, Storyline>topicToStoryLinesRDD using mapToPair transform by applying rules and dictionaries by topic to storylines
    end for
else if technique used is feature selection then
    {Categorize storylines under feature selection events}
    Build RDD fsStoryLinesRDD using map transform by applying rules and events to storylines
else if technique is svm then
    {Categorize storylines under svm events}
    Build RDD classifierStoryLinesRDD using map and filter transforms by applying rules and scored storylineVectorRDD against model
    if score $\geq$ threshold and match rules then
        assign storyline to event
    end if
end if
```

Algorithm 4. Generate Prospects.

```
Input: $\{storyline_i\}$,$\{labeled_j\}$ {storylines and labeled storylines for supervised learning}
Output: $\{prospect_k\}$
    {top scoring storylines against model in supervised learning or in unsupervised learning and the data elements they were built from}
{step 1: Use dictionaries generated in previous algorithm}
Get labeledVectorsRDD from previous step
{step 2: Use output from applied algorithm from previous step to score storylines}
if technique used is topic modeling then
    {match entity in each topic with storylines entity}
    for all topic $\in$ Topics do
        {Generate top scoring storylines by topic based on topic weights of entities}
        Score storylines entities against topic entity and sum weights of entities in topic
        Build list of top $K$ storylines by topic
    end for
else if technique is feature selection based on information gain then
    {Generate top scoring storylines with top info gain entities}
    score each storyline as a combination of its entities info gain weight
else if technique is classifier based then
    {Generate top scoring positively labeled storylines }
    Get entities of top scoring storylines against classifier
end if
{step 3: Generate list of tweets from entities of top storylines and get most frequent people and organization from tweets}
Lookup tweets associated with the top $M$ storyline entities
PairRDD <String, Integer>as count of people and organizations from tweets of top storyline entities
```

- Determine Spatial and Temporal terms: This module determines the spatial and temporal terms using the GATE API [8]. Each storyline is broken down into entities in parallel and in each process GATE APIs are initiated and used to label entities in the storyline document. The entities identified in processing step are used to create an index of entity strings to integers that is then used on storyline vectors in subsequent step.

- Build Themes and theme dictionary terms: The vectors built in processing step for each storyline are passed to one of the three theme generation routines.
  1. Topic Modeling based: When theme building process specified is topic modeling the vectors are passed to the MLLib LDA based topic modeling technique [2]. This technique returns the entities for the topics and their corresponding weights for the topic. These are then saved as dictionary for the theme.
  2. Entity Selection based: If the specified theme building process is entity selection based on information gain, the vectors are passed into the information gain based entity selection routine based on Maximum Relevancy Minimum Redundancy [5]. This technique performs information gain in parallel to generate a list of top k entities for the labeled training set. This list is saved as the dictionary for the theme event on which the labels in training data are based.
  3. Classification based: For chosen theme building process classification, the labeled data for storylines is used to build an SVM model using the MLLib SVM Spark routine [3] that build the model in parallel. This model is then used to score the storylines and top k positively labeled storylines entities are chosen and added to the dictionary.

4.2 Events Creation and Storyline Categorization

These modules assign storylines to generate events in a scalable way. Storylines event assignment module scores each storyline and determines which event they will be assigned to based on user specified rules.

- Generate Events: This module generates events as a combination of themes and spatiotemporal entities.
  1. Topic Modeling: In case of topic modeling an event is generated for each topic with the top theme entity of the topic, top location entity and top time entity by weight combined to generate the spatio-temporal event.
  2. Feature Selection: In case of feature selection top weighing theme, location and time entity with highest information gain value are combined to generate the event that corresponds to class of labeled data.
  3. Classifier: In case of classifier, the most frequent theme, location and time in positively labeled training data is combined to generate event.
- Test storylines: This job loads theme, location and time dictionaries in cache and is used to test storylines in parallel to identify storylines that fall within the event. Rules provided by end user are broadcast to all nodes and processed and information in the rules is used to determine how to assign storyline to event.
- Categorize storylines into events: This job categorizes and lists in parallel events and the storylines within them. Events are a combination of theme, location and time in format *theme:location:time*.

4.3 Prospect Identification

These modules identifies top storylines and computes the tweets associated with the entities in those storylines. It then applies named entity recognition using GATE on the text of the tweets to generate the top people and organizations. Storylines are scored in each of the techniques using the weight of entities in topic, the score assigned to storyline by classifier and the weight associated with entities in MRMR feature selection.

- Generate Top Storylines: This module generates top storylines from SVM scores in case of classification, entity weight in topic in topic modeling and feature weight in feature selection.
  1. Topic Modeling: In case of topic modeling a storyline is associated with a score that is the sum of the entity weights for each topic.
  2. Feature Selection: In case of feature selection storyline is given a score which is the sum of the score of the entities of storyline.
  3. Classifier: In case of classifier, the storyline is scored against SVM model and is assigned a score.
- Collect tweets associated with storyline entities: This job generates a list of tweets associated with each entity of a storyline.
- Extract named entities from tweets as people and organizations: This job generates a list of people and organizations that are in the list of tweets associated with top scoring storylines by technique and lists them by frequency.

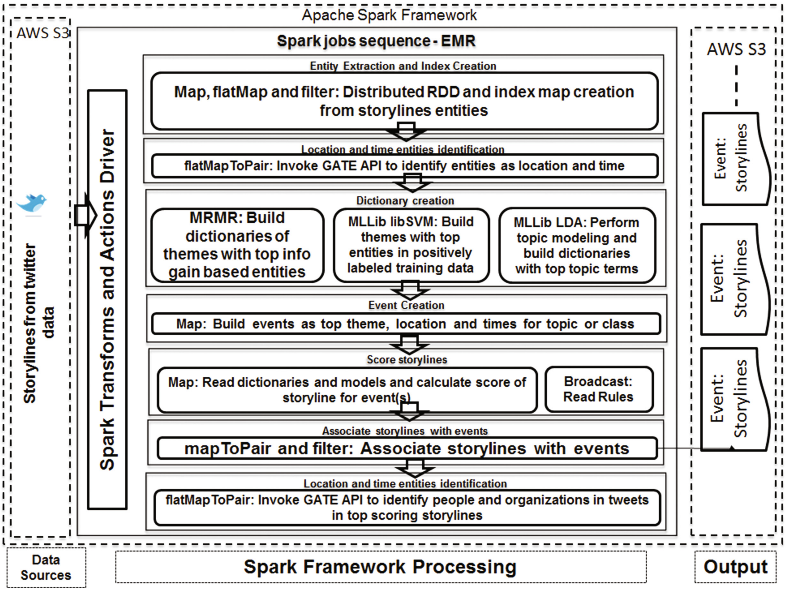
Apache Spark Framework
AWS S3
Spark jobs sequence - EMR
Entity Extraction and Index Creation
Map, flatMap and filter: Distributed RDD and index map creation from storylines entities
Location and time entities identification
flatMapToPair: Invoke GATE API to identify entities as location and time
Dictionary creation
MRMR: Build dictionaries of themes with top info gain based entities
MLLib libSVM: Build themes with top entities in positively labeled training data
MLLib LDA: Perform topic modeling and build dictionaries with top topic terms
Event Creation
Map: Build events as top theme, location and times for topic or class
Score storylines
Map: Read dictionaries and models and calculate score of storyline for event(s)
Broadcast: Read Rules
Associate storylines with events
mapToPair and filter: Associate storylines with events
Location and time entities identification
flatMapToPair: Invoke GATE API to identify people and organizations in tweets in top scoring storylines
Storylines from twitter data
Spark Transforms and Actions Driver
AWS S3
Event: Storylines
Event: Storylines
Event: Storylines
Data Sources
Spark Framework Processing
Output

Fig. 4. DISTL architecture.

5 Use Cases

This section describes the use cases for event categorization and finding prospects in various distinct domains. Section 5.1 describes the event categorization use cases while Sect. 5.2 illustrates finding top prospects through DISTL. Section 5.3 presents performance evaluation results.

5.1 Event Categorization Use Cases

Extensive experiments were performed to validate event creation and storylines categorization as scalable and useful to analysts. Two sets of storylines built on Twitter data extracted using keywords filtering using Twitter streaming API in June 2015 were used. This section describes events generated for commodity oil and events on currency Euro.

The topic modeling technique uses LDA and produces topics with top keywords and weights added to topic's dictionary. After experiment with several topic numbers, the subject expert decided on 3 as optimal number of topics from which to generate events. That number was within the constraints of not subdividing an event into subevents and yet generate meaningful results. Feature Selection based on Information gain generates a dictionary of themes with highest information gain. The features selection routine is given a set of labeled observations based on a specified class. The training data set is generated by a subject matter expert. For oil dataset a training set was built for storylines related to oil prices and for euro dataset the training data was for Greek exit from Euro. The classifier based themes dictionary creation uses labeled data provided by subject matter expert. It uses the training set to build a classifier. The most frequent features are used as dictionary and top theme, location and time terms are used to create the event. Storylines with score higher than a threshold as classified by model and entities satisfying the categorization rules are associated with the event. Storylines in each event are assigned by testing the entities against event's location, time and theme entities. The rule used to categorize storylines into events in our experiments is "location *or* time *or* theme".

Oil Events. This subject is regarding tweets related to the commodity oil. The filtering keywords for tweets extracted are 'oil, wti, brent, crude, frack, fracking, oil price, petroleum, petro, west texas intermediate, oil inventory, oil production, oil refinery, #crudeoil and nymex'. The entry points for storylines are 'oil' and 'petroleum'.

Applying topic modeling produces the topics shown in Table 1. The top entities for each topic are similar, two 'amp' and 'petroleum' being top entities in each of the 3 topics and the third ranked entities being number 24, gas and canvas indicating one of the topics is more related to painting oil. Applying information gain based feature selection generates the top terms shown in Table 1. These weights more accurately reflect oil related storylines as expected of a supervised technique as the top information gain weight terms are not only related to oil but

Table 1. Oil topic modeling, info gain based feature selection and classifier labeled data based top entity weights.

Topic1	Weight	Topic2	Weight	Topic3	Weight
amp	0.033	amp	0.32	amp	0.0334
petroleum	0.023	petroleum	0.0266	petroleum	0.023
24	0.0150	gas	0.0129	canvas	0.011
Feature	Info gain	Feature	Info gain	Feature	Info gain
US $61.53	0.0335	weather fueling oil price discovery	0.0267	global crude oil price	0.0294
Singapore	0.006	oil prices	0.0288	17-Jun	0.0091
Feature	Weight	Feature	Weight	Feature	Weight
petroleum	333	US $61.53	58	oil price	77
London	44	reuters	61	17-Jun	15

also to the price of oil which was the basis of labeled data. Top features by classifier are in Table 1. These features accurately reflect the key entities in training dataset specifically the ones most frequent in positively labeled elements.

Top storylines based on events from top feature selection entities are in Table 2. Top storylines by weight for each topic from topic modeling as generated by application of rules are in Table 2. Top storylines based on events from classifier based top themes, locations and time entities are in Table 2. These categorizations clearly show that storylines in same set get categorized differently under different events simply due to application of different entity ranking techniques.

The analysis of oil brought mixed results, which is potentially explained by the broad range of topics covered by the term oil. For example, oil is associated with petroleum, body oil, oil paintings and other categories unrelated to the area of focus. Thus, the topics and associated storylines identified by topic modeling shown in Tables 1 and 2 hold limited relevance and add little to the understanding of oil prices. The results from topic modeling were also clouded by the entity 'amp' which actually refers to the character '&' and was erroneously picked up as an entity. However, the information gain and classification models did reveal interesting topics and storylines. First, the international crude oil price of Indian Basket as computed by the Ministry of Petroleum and Natural Gas of the Republic of India was $61.53 per barrel on June 16, 2015, an informative metric given India is one of the largest importers of petroleum in the world and the Indian government heavily subsidizes those imports. Second, the entity 'weather fueling oil price discovery' alluded to the foul weather moving through Texas at that time which was expected to impact oil production and thus prices.

Euro Events. This subject involves tweets regarding the change in value and status of the currency euro. Filtering keywords 'Euro, Euro exchange rate, euro

Table 2. Oil top storylines by LDA, Info Gain and classifier weight.

Topic event	Storyline	LDA weight
amp:north america:1988	oil:london #jobs:amp:sales manager	0.0422
amp:north america:1988	oil:amp:deals:funnel	0.0400
petroleum:london:17-Jun	oil:london #jobs:amp:sales manager	0.0422
petroleum:london:17-Jun	amp:seed oil unrefined 100:deals	0.0401
gas:greece:today	oil:amp:engine oil:gas	0.0480
gas:greece:today	oil:grease:paper 10 sets face care:amp	0.0470
Topic event	Storyline	Info gain weight
US $61.53:singapore:17-Jun	oil:petroleum:us $61.53:global crude oil price	0.0629
US$61.53:singapore:17-Jun	oil:petroleum:us $61.53:weather fueling oil price discovery	0.0602
US $61.53:singapore:17-Jun	oil:reuters:oil prices:production	0.0513
Topic event	Storyline	Classifier weight
petroleum:london:17-Jun	oil:long:iran:petroleum	1.2209
petroleum:london:17-Jun	oil:petroleum:oil price:our 2015 global #oil	1.1375
petroleum:london:17-Jun	oil:brent oil forecast. current price:petroleum:global crude oil price	1.1114

rates, euro-dollar, dollar euro, euro crisis, euro conversion, euro rate, eur and eur usd' are specified for tweet collection. The entry point for storylines is 'euro'.

On euro related storylines we applied event generation techniques. Applying topic modeling generates the topics shown in Table 3. Three topics were provided to the LDA method. The top entities were similar for the three topics with the difference being in third highest weighted entity indicating topics being related to emergency summit over Greek crisis. The top features by classifier are shown in Table 3. These features are more accurately related to the Greek exit due to the application of training data on the subject provided. Applying information gain based feature selection produces the top terms shown in Table 3. These entities are also highly relevant due to use of training data. Top storylines by weight for each topic as generated by application of rules are in Table 4. Top storylines based on events from top feature selection entities are given in Table 4. Top storylines based on events from classifier based top themes, locations and time entities are provided in Table 4. These storylines clearly show the preponderance of storylines on Greek exit crisis from the Euro at the time and the Federal Open Market Committee meeting on June 18, 2015.

On the analysis of the Euro dataset, topic modeling, information gain, and classification all highlighted the crisis occurring in Greece's economy and the potential of a Greek Exit from the Euro. Topic modeling even highlighted the emergency summit taking place in Luxembourg to discuss the situation. In this case, the information gain based feature selection analysis generated the most noise as the highest weighted features included indiscernible numbers and entities related to sports even though two of the features were 'greek exit' and 'syriza

Table 3. Euro top entities by info gain feature selection, topic modeling and classifier weights.

Topic1	Weight	Topic2	Weight	Topic3	Weight
The euro	0.0136	the euro	0.0137	the euro	0.0139
eur	0.012	eur	0.0123	eur	0.0121
Luxemborg	0.0055	emergency summit	0.0064	2015	0.008
Feature	Info gain	Feature	Info gain	Feature	Info gain
Greek exit	0.0621	0.049	0.0302	0.03	0.0461
18 june #football #soccer #sports	0.0282	0.08	0.0473	syriza hardliners back	0.0228
Feature	Weight	Feature	Weight	Feature	Weight
Greek exit	194	yesterday's fomc meeting	91	72.43	108
2015	24	1199.9	97	greece	83

hard-liners back'. The number of storylines an analyst has to review is greatly reduced for events, for SVM the number of storylines is reduced to 933 from over 300000 when threshold of 1.0 is set for the SVM scores.

5.2 Identify Key Prospects Use Cases

DISTL provides users with the ability to track investment and target prospects by identifying emerging people and organizations in open data. Further, DISTL narrows down the number of prospects down to the ones identified in the storylines with highest scores based on the models built to identify them. Two sets of storylines built on Twitter data extracted using keywords filtering using Twitter streaming API in July 2016 were used. Section 5.2 provides details of identifying top prospects within Autonomous Car domain and Sect. 5.2 provides details on top prospects for Augmented Reality/Virtual Reality (AR/VR) domain.

Autonomous Car Investments. This use case identifies top people and organizations with the highest impact in domains of connected car and autonomous driving. Filtering keywords 'Autonomous vehicle, Mobile Car Sharing, Telematics, driverless vehicle, ..., driver-assistance systems' are specified for tweet collection. The entry points for storylines are 'connected car, driverless car'.

On connected car related storylines we applied the three supervised and unsupervised techniques we had earlier used in event generation. Applying topic modeling generates the topics and entity weights as shown in Table 5 and corresponding storyline weights shown in Table 6. Three topics were provided to the LDA method. The top storylines by classifier weight are shown in Table 6. Applying information gain based feature selection produces the top entities in Table 5

Table 4. Euro top storylines by event and topic modeling, info gain and classifier weights.

Topic event	Storyline	LDA weight
the euro:2015:luxemborg	euro:zone ecofin meetings:the euro:eur	0.03106
the euro:2015:luxemborg	euro:dibebani yunani:eur:the euro	0.02900
eur:19-Jun:greece	euro:zone ecofin meetings:the euro:eur	0.0309
eur:19-Jun:greece	euro:dibebani yunani:eur:the euro	0.0288
amp:this day:edinburg	euro:zone ecofin meetings:the euro:1.7:0	0.0207
amp:this day:edinburg	euro:lows:6:eur	0.0215
Topic event	Storyline	Info gain weight
greek exit:greece:18 June	euro:2:0.08:#dollar	0.0068
greek exit:greece:18 June	euro:gold:0.13:0.08	0.029
greek exit:greece:18 June	oil:gas temp:marks sattin:#cash #applications accountant	0.29
Topic event	Storyline	Classifier weight
greek exit:greece:2015	euro:yesterday's fomc meeting:greek exit:support	1.3681
greek exit:greece:2015	euro:yesterday's fomc meeting:greek exit:greece #euro	1.3410
greek exit:greece:2015	euro:central bank:greeks themselves:greece	1.2701

Table 5. Connected car top entities by info gain feature selection, topic modeling and classifier frequency weights.

Topic1	Weight	Topic2	Weight	Topic3	Weight
connected car	0.1659	connected car	0.1626	connected car	0.1586
driverless car	0.0883	driverless car	0.0863	driverless car	0.0900
navigation	0.0597	Chicago	0.0542	navigation	0.0628
Feature	Info gain	Feature	Info gain	Feature	Info gain
1	0.0827	car	0.0327	venturecapital	0.0106
wingz	0.005	chicago	0.0071	0	0.0120
Feature	Frequency	Feature	Frequency	Feature	Frequency
China	194	alibaba	91	volkswagen	108
internet	180	reuters	97	safety	83

and top storylines shown in Table 6. These entities are also highly relevant due to use of training data.

Top storylines by weight for each topic as generated by application of rules are in Table 6. Top named entities in data elements for top storylines are shown in Table 7.

Table 6. Connected car top storylines by topic modeling, info gain and classifier weights.

Storyline	LDA weight	Topic
connected car:automotive:car platform:technews	0.2017	topic 1
connected car:lg:technology:the road	0.2010	topic 1
connected car:china:economy:new post	0.2021	topic 2
connected car:alibaba's internet:car:new post	0.2020	topic 2
connected car:jobs:northbrook:systems design	0.2022	topic 3
connected car:alibaba's internet:car:the world	0.2019	topic 3
Storyline	Info gain weight	
connected car:car:nokia growth partners:unquotenews	0.1162	
connected car:car:jaguar land rover's new tech subsidiary inmotion:smartcommuter	0.1051	
connected car:china:navigation:the navyo smart glove	0.0870	
Storyline	Classifier weight	
connected car:alibaba's internet:car:connected	0.1167	
connected car:automotive:programs:mapping	0.1154	
connected car:automotive:the google self:mapping	0.1145	

Results of the connected car analyses proved useful in terms of understanding the key connected car events of the day as well as identifying the individuals and organizations related to those events. Each algorithm produced different, yet complementary results that provided the analyst with a broad picture of the different themes emanating from the day's Twitter data. For instance, topic modeling highlighted LG's new partnership with Volkswagen to develop technology to allow drivers to monitor and control devices in the home from their car as well as improve in-vehicle entertainment capabilities. Topic modeling also highlighted the newly developed car with the YunOS operating system jointly developed by China's Alibaba's and SAIC Motors. The information gain algorithm produced different results that emphasized Jaguar Land Rover's new venture in the area car sharing through their wholly owned subsidiary InMotion and Nokia Growth Partner's new internet-of-things investment fund, which invests in connected car companies. The classification results partially overlapped with the topic modeling results, as both identified Alibaba's activities. And classification added additional value by also highlighting updates to Google's self-driving car initiative, which was not captured in top storylines from topic modeling. Ultimately, the list of storylines identified by the three algorithms was distilled to a list of persons and organizations found within the data underlying those storylines (Table 4). While some organizations were mischaracterized as persons, including ZF Friedrichshafen (a German auto-parts manufacturer) and BABA (the ticker symbol for Alibaba), the algorithm did an adequate job of identifying

Table 7. Connected car top named entities by info gain feature selection, topic modeling and classifier weights.

Organization	Topic frequency	Person	Topic frequency
LG	40	ZF Friedrichshafen	1
VW	9	Lee Coleman	1
Ford	2	ena	1
Organization	Info gain frequency	Person	Info gain frequency
LG	68	aren	3
Volkswagen	42	ena	1
Reuters	20	BABA	1
Organization	Classifier frequency	Person	Classifier frequency
LG	40	Ethan Lou	1
Reuters	20	Lee Colman	1
Tesla	4	aren	1
Google	4	Jake Spring	1

relevant and interesting entities. Tesla and Google's efforts in connected car are well known, but LG's initiatives were certainly less publicized. The individuals identified included connected car expert Lee Colman, Head of Connected Car at SBD Automotive (a UK based consultancy). Mr. Colman was also a frequent speaker at various conferences including CarConExpo Dusseldorf 2016 and Telematics Berlin 2016 and is, perhaps, an expert worth following. Other individuals included Ethan Lou and Jake Spring, both Reuters reporters who cover the car industry.

Augmented Reality/Virtual Reality Prospects. This use case identifies top people and organizations that are having the highest impact in nascent augmented and virtual reality domains. Filtering keywords 'virtual reality, augmented reality, ..., vr bandwidth, ar bandwidth, vr optimization, ar optimization, amazon vr, amazon ar' are specified for tweet collection. The entry points for storylines are 'vr, ar, virtual reality'.

On augmented reality related storylines we applied the three supervised and unsupervised techniques we had earlier used in event generation. Applying topic modeling generates the topics shown in Table 8. Three topics were provided to the LDA method. The top entities were similar for the three topics with the difference being in third highest weighted entity. The top features by classifier are shown in Table 8. Applying information gain based feature selection produces the top terms shown in Table 8. These entities are also highly relevant due to use of training data. Top storylines by weight for each topic as generated by application of rules are in Table 9. Top storylines based on events from top feature selection entities are given in Table 9. Top storylines based on events from classifier based top themes, locations and time entities are provided in

Table 8. AR top entities by info gain feature selection, topic modeling and classifier weights.

Topic1	Weight	Topic2	Weight	topic3	Weight
Virtual reality	0.0901	Virtual reality	0.1307	vr	0.1281
The future	0.0264	The future	0.0901	Virtual reality	0.0912
The magic leap	0.0260	Startup	0.0260	Startup	0.0297
Feature	Info gain	Feature	Info gain	Feature	Info gain
Virtual reality	0.0215	0	0.1083	1	0.0549
2016	0.005	vr	0.0207	Our blog	0.0110
Feature	Frequency	Feature	Frequency	Feature	Frequency
Games	58	Oculus	61	News	18
Reality	175	Headset	99	Pokemon	24

Table 9. AR top storylines by event and topic modeling, info gain and classifier weights.

Storyline	LDA weight	Topic
vr:cool:front:i recalibrate the oculus sensor	0.2031	topic 1
vr:cool:la france:metheniyacine	0.2023	topic 1
vr:google:reality #technology:tango:vr	0.2044	topic 2
vr:htc vive:people:robbers	0.2038	topic 2
vr:amazing:oculus:verder gaan	0.2021	topic 3
virtual reality:samsung:tech:three companies	0.2020	topic 3
Storyline	Info gain weight	
vr:eyes:zuckerberg:so boring	0.1462	
vr:the australian public's attitude:exploring:yuri_librarian it's cool	0.1462	
vr:atari cofounder nolan bushnell:startups:virtual reality	0.1103	
Storyline	Classifier weight	
vr:la france:le grand cran:vive	2.0841	
vr:1:11:us a mobile monday detroit	1.9923	
vr:1:2016:titres playstation vr annoncés	1.980	

Table 9. Top named entities in the data elements for top storylines are shown in Table 10.

Besides well known companies in Table 10, the bulk of named entity list comprises of new companies such as 'Unimersiv - VR Education', 'Lunar Flight CV' and 'opTic' that are of specific interest to investors. The utility of the virtual reality and augmented reality results were more mixed compared to the connected car results. Themes resulting from topic modeling varied from Lenovo's use of Google's Tango augmented reality technology in their new smartphones to robbers using the mobile game Pokemon Go to find and lure unsuspecting

Table 10. AR top named entities by info gain feature selection, topic modeling and classifier weights.

Organization	Topic frequency	Person	Topic frequency
HTC Vive	40	VR	46
Sony	12	Unity	10
Microsoft	13	David	4
Organization	Info gain frequency	Person	Info gain frequency
HTC Vive	28	des	9
Sony	12	phil	6
Microsoft	9	lucas	7
Organization	Classifier frequency	Person	Classifier frequency
HTC Vive	47	VR Shell	25
Google	45	Kevin Durant	7
Samsung	35	vida	7
Nintendo	9	una	9

victims, and the impending launch of Samsung Gear VR 2, Samsung's latest virtual reality headset. Storylines identified thru information gain and classification included publicity for mobile Monday Detroit, an event showcasing AR and VR demos, Mark Zuckerberg's announcement of Open Cellular which is tangentially related to the topic of virtual reality, and news around the development of Sony's Playstation VR gaming platform. The organizations identified by the algorithms included some of the biggest names in virtual reality such as Google, HTC, Sony, and Samsung as well as other organizations less known for their efforts including Nintendo and Microsoft. Nintendo was making a huge splash in the market with its wildly successful Pokemon Go augmented reality mobile game and this analysis perhaps acted as an early indicator of how successful the game would eventually become. Meanwhile, Microsoft was making news for using virtual reality for autistic kids in their preparation for jobs. The identified persons included Phil (Phil Spencer) the Xbox chief at Microsoft, David (David Joyner) of Georgia Tech who publishes a blog summarizing news and events in the area of virtual reality, and Kevin Durant, professional basketball player, who was making headlines as the Golden State Warriors used virtual reality to recruit Durant. As was the case with connected car, some of the identified names were actually organizations or other entity types. One such example was Unity, a game development platform used by leading game developers. Overall, the algorithms effectively managed to refine a large volume of raw data (over 140,000 virtual reality related tweets and nearly 40,000 connected car related tweets) to a much more manageable and targeted dataset for the analyst to examine and further research. Furthermore, it would be interesting to track the results of multiple days' worth of data to see what results would emerge.

5.3 Performance

The performance of the techniques used in event creation at different levels of distribution is evaluated in this subsection. The results for running the techniques on various sized clusters are presented. The experiments were run on AWS using Elastic MapReduce clusters running Spark. This allows for clusters to be configured on demand on the cloud so that scalability of the techniques on different sized datasets and clusters can be tested. Cluster nodes are of type m3.2xlarge with 8 vCPU processors and 30 GB of RAM.

Figure 5 shows the performance of topic modeling on various sized clusters. The same code run on a single node is an approximation of how similarly written single node sequential version will perform. The results show clearly that with increasing number of storylines, the time taken to perform topic modeling on the storylines does not increase significantly on an 8 node cluster but continues to increase for sequential runs. Beyond a dataset of certain size the single node execution generates out of memory errors. Topic modeling is highly iterative hence its distribution is critical to its being able to scale to larger datasets. Results are similar for Information gain based features selection and SVM modeling executions on multiple sized training datasets.

Figure 6 shows the performance of spatio-temporal entity identification. The results clearly show that the process of identifying spatial and temporal entities is highly parallelizable with testing each storyline against GATE API independent of others. Figure 8 shows storylines categorization performance using feature selection generated information gain weights. These results show that once feature selection has generated top info gain entities, categorizing storylines under those events is highly parallelizable and scalable with running times staying stable with increasing data and cluster sizes. Figure 7 shows results for categorizing storylines into events using topic modeling weights. This was done for 3 topics

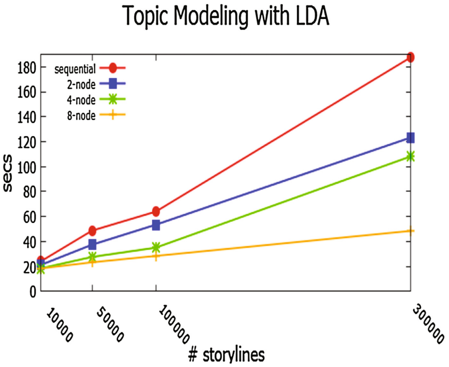

Fig. 5. Performance of topic modeling on various cluster and storyline data sizes.

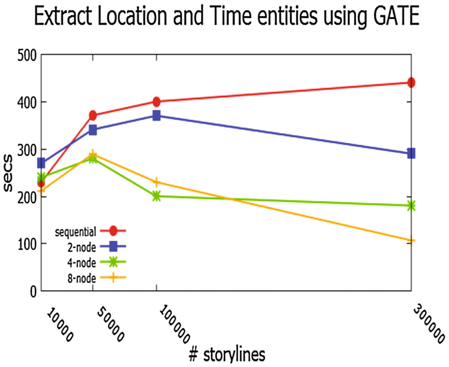

Fig. 6. Performance of spatio-temporal entity creation on various cluster and storyline data sizes

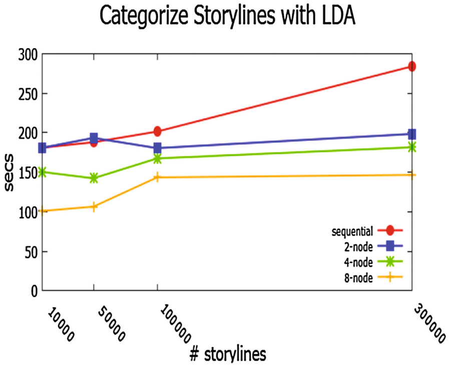

Fig. 7. Performance of storylines categorization into events generated from topic modeling on various cluster and storyline data sizes

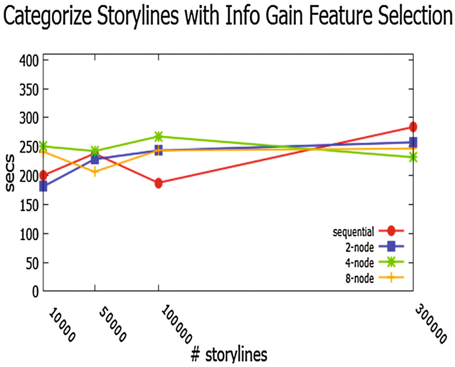

Fig. 8. Performance of storylines categorization into events generated from feature selection on various cluster and storyline data sizes

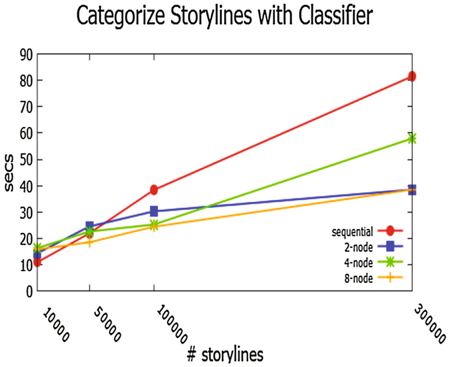

Fig. 9. Performance of storylines categorization based on svm model scoring on various cluster and storyline data sizes.

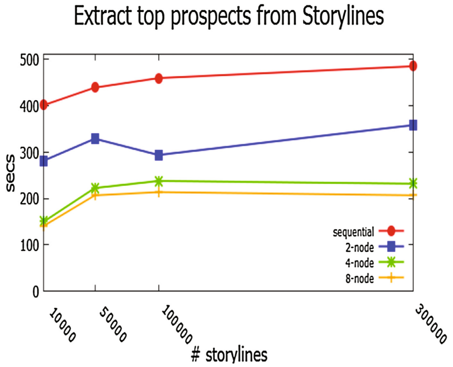

Fig. 10. Performance of extracting top people and organizations from tweets of top storylines as identified by FS, LDA and SVM on various cluster and storyline data sizes.

and each storyline was tested against multiple events, yet the process is highly scalable and parallelizable.

Figure 9 shows the performance of scoring storylines using SVM classifier. This is also highly parallelizable as once a model is built, it can score storylines independent of each other. Figure 10 shows the performance of top named entities as organizations and people extraction to identify top prospects from tweets of top storylines. The performance over increasing larger test datasets and clusters shows that this is highly scalable.

The results clearly show the scaling of events generation for storylines for large datasets. Increasing the size of cluster allows full horizontal scaling in DISTL. Increased overhead of Spark in some cases results in deterioration in performance on small clusters as compared to serial execution on small datasets but with larger sets of storylines the performance improves vastly.

6 Conclusions

Event based storylines categorization and prospect identification are critical in providing analysts meaningful ways to analyze large number of storylines generated from social and open data. Proposed framework DISTL is effective in generating events and identifying prospects at a large scale. The supervised and unsupervised learning techniques incorporate domain expert knowledge and generate events and prospects from storylines. Rules allow flexibility in categorizing storylines under events. The resulting events incorporate theme, location and time and are useful in allowing analyst analyze large number of storylines. Experiments show that the validity of results in multiple domains and effectiveness of in-memory techniques to scale to large amounts of data by simply increasing the cluster sizes as needed.

References

1. Agarwal, C., Subbian, K.: Event detection in social streams. In: SDM, pp. 624–635 (2012)
2. Apache and Spark (2015a). https://spark.apache.org/docs/latest/mllib-clustering.html#latent-dirichlet-allocation-lda
3. Apache and Spark (2015b). https://spark.apache.org/docs/latest/mllib-linear-methods.html\#inear-support-vector-machines-svms
4. Apache and Spark. Spark programming guide (2015c). http://spark.apache.org/docs/latest/programming-guide.html
5. Apache, Spark, and Packages (2015). https://github.com/wxhc3sc6opm8m1hxbomy/spark-mrmr-feature-selection
6. Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D., Ertl, T.: Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 143–152 (2012)
7. Chen, F., Neill, D.B.: Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graph. In: ACM SIGKDD, pp. 1166–1175 (2014)
8. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: Developing language processing components with gate version 8. University of Sheffield Department of Computer Science (2014)
9. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)

10. Weng, J., Lee, B.-S.: Event detection in twitter, pp. 401–408. AAAI (2011)
11. Keskisärkkä, R., Blomqvist, E.: Semantic complex event processing for social media monitoring-a survey. In: Proceedings of Social Media and Linked Data for Emergency Response (SMILE) Co-Located with the 10th Extended Semantic Web Conference, Montpellier, France, CEUR Workshop Proceedings (2013)
12. Lappas, T., Vieira, M.R., Gunopulos, D., Tsotras, V.J.: On the spatiotemporal burstiness of terms. Proc. VLDB Endow. **5**(9), 836–847 (2012)
13. Leetaru, K., Schrodt, P.A.: GDELT: global database of events, language, and tone. In: ISA Annual Convention (2013)
14. Li, C., Sun, A., Datta, A.: Twevent: segment-based event detection from tweets. In: Conference on Information and Knowledge Management, pp. 155–164 (2012a)
15. Li, R., Lei, K.H., Khadiwala, R., Chang, K.: Tedas: a twitter-based event detection and analysis system. In: Proceedings of 28th IEEE Conference on Data Engineering (ICDE), pp. 1273–1276 (2012b)
16. Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., Shrimpton, L.: Can twitter replace newswire for breaking news? In: 7th International AAAI Conference on Weblogs and Social Media (ICWSM) (2013)
17. Pohl, D., Bouchachia, A., Hellwagner, H.: Automatic sub-event detection in emergency management using social media. In: Proceedings of the 21st International Conference Companion on World Wide Web (WWW 2012 Companion), pp. 683–686, New York, NY, USA. ACM (2012)
18. Radinsky, K., Horvitz, E.: Mining the web to predict future events. In: WSDM, pp. 255–264 (2013)
19. Reuter, T., Buza, L.D.K., Schmidt-Thieme, L.: Scalable event-based clustering of social media via record linkage techniques. In: ICWSM (2011)
20. Saha, A., Sindhwani, V.: Learning evolving and emerging topics in social media: a dynamic NMF approach with temporal regularization. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM 2012), pp. 693–702, New York, NY, USA. ACM (2012)
21. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: WWW, pp. 851–860 (2010)
22. Shukla, M., Santos, R.D., Chen, F., Lu, C.-T.: Discrn: a distributed storytelling framework for intelligence analysis. Virginia Tech Computer Science Technical report (2015)
23. Walther, M., Kaisser, M.: Geo-spatial event detection in the twitter stream. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 356–367. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36973-5_30
24. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Presented as Part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2012), pp. 15–28, San Jose, CA. USENIX (2012)
25. Zhao, L., Chen, F., Dai, J., Lu, C.-T., Ramakrishnan, N.: Unsupervised spatial events detection in targeted domains with applications to civil unrest modeling. PLOS One **9**(10), e110206 (2014)
26. Zhao, L., Chen, F., Lu, C.-T., Ramakishnan, N.: Spatiotemporal event forecasting in social media. In: SDM, pp. 963–971 (2015)
27. Zhou, X., Chen, L.: Event detection over twitter social media streams. VLDB J. **23**(3), 381–400 (2014)

28. Li, J., Tsang, E.P.K.: Investment decision making using FGP: a case study. In: Proceedings of the 1999 Congress on Evolutionary Computation (CEC 1999), vol. 2, no. 2, pp. 1259–1279 (1999)
29. Shukla, M., Santos, R.D., Fong, A., Lu, C.-T.: DISTL: distributed in-memory spatio-temporal event-based storyline categorization platform in social media. In: Proceedings of the 2nd International Conference on Geographical Information Systems Theory, Applications and Management, pp. 39–50, Italy, Rome (2016)
30. Cheh, J.J., Weinberg, R.S., Yook, K.C.: An application of an artificial neural network investment system to predict takeover targets. J. Appl. Bus. Res. (JABR) **15**(4), 33–46 (2013)
31. Geum, Y., Lee, S., Yoon, B., Park, Y.: Identifying and evaluating strategic partners for collaborative R&D: index-based approach using patents and publications. Technovation **33**(6), 211–224 (2013)