SocialCube: A Text Cube Framework for Analyzing Social Media Data

Xiong Liu†, Kaizhi Tang†, Jeffrey Hancock‡, Jiawei Han#,
Mitchell Song†, Roger Xu†, Vikram Manikonda†, Bob Pokorny†
†Intelligent Automation, Inc.
‡Department of Communication, Department of Information Science, Cornell University
#Department of Computer Science, University of Illinois at Urbana-Champaign
xliu09@gmail.com, {ktang, msong, hgxu, vikram, bpokorny}@i-a-i.com,
jeff.hancock@cornell.edu, hanj@illinois.edu

***Abstract*—The recent development of social media (e.g., Twitter, Facebook, blogs, etc.) provides an unprecedented opportunity to study human social cultural behaviors. These data sources provide rich structured data (e.g., XML, relational tables, and categorical data) as well as unstructured data (e.g., texts). A significant challenge is to summarize and navigate structured data together with unstructured text data for efficient query and analysis. In this paper we introduce a text cube architecture designed to organize social media data in multiple dimensions and hierarchies for efficient information query and visualization from multiple perspectives. For example, an affective process cube allows the analyst to examine public reaction (e.g., sadness, anger) to a range of social phenomena. The text cube architecture also supports the development of prediction models using the summarized statistics stored in a data cube. For example, models that detect events, such as violent protests in the Egyptian Revolution, can be built using the linguistic features stored in an event data cube. These kinds of models represent higher level of knowledge representation and may help to develop more effective strategies for decision-making based on social media data.**

Keywords-Tex cube; social media; language processing; feature analysis; text mining; data mining; human social cultural behavior

I. INTRODUCTION

The exponential growth of text-based communication associated with the Internet has led to a vast increase in the amount of social media and unstructured text that are not currently warehoused or mined in ways relevant to Human Social Cultural Behavior (HSCB) analysis. HSCB analysis and modeling is an emerging research area that focuses on understanding, predicting, and shaping human behaviors cross-culturally [1]. Current systems analyzing textual HSCB data typically do not take advantage of the state-of-the-art in automated language analysis. Also, human behavior is continuously changing, and as a result, HSCB data can quickly become out of date and not suitable for dynamic HSCB analysis. A critical task is to develop a system that can dynamically collect and warehouse unfiltered textual communication data and make available the data to state-of-the-art automated linguistic analysis for HSCB modeling and applications. In building this system, we encounter the following challenges:

- The HSCB data exhibit **complex features**. The data are multi-dimensional, time stamped, geospatially referenced, from multiple sources, in multiple data types, and associated with semantic tags. More data with different features are being added regularly. How to develop a stable but extendible architecture to integrate data with complex features is a challenging problem.
- For information consumers to access and query the massive amounts of HSCB data, **efficient data warehousing and mining capabilities** must be developed to prepare data summation and knowledge for fast information access. How to build effective and time critical data cubing and optimized knowledge discovery methods are both challenging.

To address these challenges, we propose a dynamic data cubing and mining system, called *SocialCube*, for large amounts of HSCB data. SocialCube is an advanced data cube architecture that allows analysts to summarize and navigate structured data together with unstructured text for efficient query and analysis. In data warehousing, data cube is a way to organize data in multiple dimensions and multiple hierarchies for information query and visualization from multiple perspectives [2]. A data cube allows data to be aggregated and viewed from multiple perspectives, and it is defined by measures and dimensions. The measures (or facts) are numeric values that are usually additive (e.g., sales of a product). Analysts need to look at measures using some "by" conditions. The "by" conditions are dimensions. For example, in order to analyze sales volume, analysts often want to see its measure by day and by location. In this sense, dimensions are the perspectives with respect to which an analyst wants to aggregate or view measures.

Unlike a traditional data cube where measures are directly retrieved from the original databases, SocialCube provides an advanced text analytics capability for extracting HSCB measures from unstructured text streams. In addition, SocialCube supports the development of prediction models using a data mining approach. Therefore, SocialCube is a large-scale, dynamic approach for collecting, organizing and analyzing text-based communications to assess HSCB dimensions (e.g., affect, deception, group identity) for a given group and to predict current belief states and likely intended actions.

In this paper, we first describe the conceptual design and architecture of SocialCube. Then we introduce the HSCB feature analysis framework which provides complex linguistic

978-0-7695-5015-2/12 $26.00 © 2012 IEEE
DOI 10.1109/SocialInformatics.2012.87

features for HSCB measures. Next we discuss the details of the text cube architecture which allows analysts to summarize and navigate structured data (dimensions) together with unstructured text (measures) for efficient query and analysis. Next we introduce the data mining approach to support the development of prediction models using the data from the data cube. We also present some case studies to demonstrate the analysis capability of SocialCube. Finally, we conclude the paper and discuss future research.

II. THE SOCIALCUBE FRAMEWORK

To solve the problem of information summarization and querying for HSCB data, we developed the SocialCube framework. Fig. 1 shows the system architecture of SocialCube. The core of SocialCube includes: 1) a data collection component, 2) a HSCB feature analysis component, 3) a text cube component, and 4) a data mining and modeling component. The following is a brief description of these components.

The data collection component automatically extracts data of interest from various data sources such as search engine, social media, and databases. The procedure for data extraction is dependent on the format and accessibility of available data sources. Since many data sources related to HSCB are websites or web services that can be accessed on the Internet, we address the data collection methods that can obtain data from the Internet. These methods include Application Programming Interface (API) invocation via web services and data scraping via parsing web pages. For example, Twitter is a microblogging website that has been useful as a source for HSCB analysis (e.g., political sentiment analysis [3], user influence [4], and spread of news [5]). The Twitter streaming API allows applications to have real-time access to tweet objects in JSON (JavaScript Object Notation) format. Using this API, we can design code to automatically extract live tweets for a topic, transform and load them into the textual database for subsequent HSCB analysis.

The HSCB feature analysis component extracts linguistic features from text using text analytics tools. These linguistic features are the basic elements for HSCB dimensions such as affect, deception, and sense of fatalism, and for any future HSCB dimension. This analysis framework addresses the selection of linguistic features with reference to theories and psychological expectations. It also provides computational techniques (e.g., feature selection in machine learning) to extract additional linguistic features that are emergent from a specific context. Using this analysis framework, we have performed case studies on affect analysis of political events such as Libya civil war and Egyptian revolution. These case studies show that there are linguistic features in the text that are predictive of HSCB dimensions.

The data cube architecture allows analysts to summarize and navigate structured data (dimensions) together with HSCB measure from unstructured text data for efficient query and analysis. The data cube has an underlying star schema database to store the high dimensional HSCB data. The star schema has a fact table that contains the linguistic measures as well as keys to each of the related dimension tables such as location and time. With the schema defined, users are able to view the cube models and perform analyses using an Online Analytical Processing (OLAP) tool.

In addition, the data mining and modeling component provides the capability to build prediction models using the data taken from data cubes or the star schema database. In this way, the values summarized in the data cube provide a powerful filter for data mining models. The model built on the summarized statistics usually represents higher levels of knowledge representation.

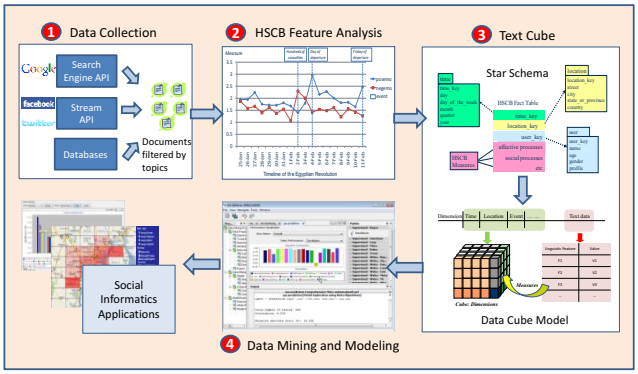

Fig. 1. SocialCube Architecture

III. HSCB FEATURE ANALYSIS OF UNSTRUCTURED TEXT

Linguistic feature analysis is a preliminary step for developing text-based data cubes and data mining methods. We have designed a comprehensive HSCB linguistic feature analysis framework that allows for an extensible set of HSCB dimensions, see Fig. 2. The framework has three layers: 1) the generic linguistic feature layer, 2) the feature selection layer, and 3) the HSCB dimension layer.

A. *Generic linguistic feature layer*

The process for developing a fully automated SocialCube system begins with the identification of linguistic features that are related to HSCB dimensions. These linguistic features may be low-level features (such as individual word-counts), high-level features (such as discourse cohesion) or anywhere on a spectrum in between. In this stage, existing text analysis tools are used to automatically generate linguistic features. Extracting low-level features may require simple word counts,

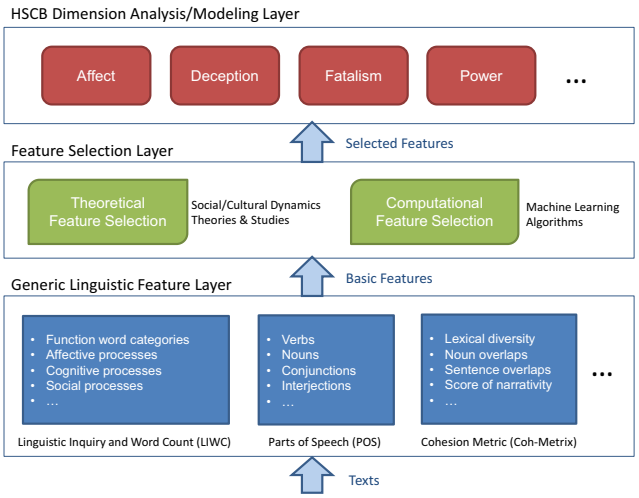

Fig. 2. Comprehensive HSCB feature analysis

while extracting higher-level features may require more sophisticated language process techniques, such as syntactic parsing and cohesion computation.

Some text analysis tools that can be used to generate linguistic features include Linguistic Inquiry and Word Count (LIWC) [6], Stanford Part-Of-Speech (POS) Tagger [7], and Coh-Metrix [8]. Together, these tools produce the basic elements for the HSCB dimensions we describe below, and for any future HSCB dimension. The key point here is that these basic elements are always being collected and tracked by the system, and higher layers can be customized by combining these basic elements into higher-level HSCB dimensions.

B. Linguistic Feature Selection layer

The SocialCube allows for the selection of linguistic features with reference to psychological and sociological expectations. These features are called theoretical (top-down) features. SocialCube also relies on computational techniques (e.g., feature selection) to extract additional linguistic features that are emergent from a specific context. These features are called computational (bottom-up) features.

1) Theoretical Features

This stage is grounded in theory and requires a deep understanding of the social dynamic under consideration, along with how the social dynamic may be manifest in discourse. Once theoretical features have been identified as potential correlates of a HSCB dimension (e.g., deception), empirical validation of these features is required. This is to establish that the theoretically predicted correlations are present and statistically significant.

For example, Newman, Pennebaker, and colleagues [9] developed an empirically-based model of deceptive language. The Newman-Pennebaker (NP) model describes four categories of words that change in relative frequency during deception: 1) fewer first person singular ("I") as liars try to distance themselves psychologically from their lie, 2) fewer exclusive words ("except", "but") as lies tend to be less complex than truthful statements, 3) more negative emotion terms that reflect the guilt and anxiety related to being deceptive, and 4) more action verbs that help move the story along and distract the listener.

The NP model has been tested in various contexts, such as detecting lies by students told in laboratory experiments [10], detecting lies by inmates in prison [11], and detecting lies by business executives in Enron emails [12]. Another important real-world context for deception is political communication.

2) Computational Features

The output for each word category from the LIWC default dictionary represents a feature in our analysis. Some of the key features from our theoretical analysis include 1st person singular (e.g., I, me), negative emotion terms (e.g., hurt, ugly, nasty), exclusives (e.g., except, but, without), and action verbs (e.g., arrive, go). We pay particular attention to these in our top-down modeling, and then add the rest of the features in our bottom-up modeling, combining them together in our final classification model.

We apply a feature selection algorithm to the LIWC outputs to extract additional linguistic features without reference to psychological expectations. This algorithm evaluates the worth of a feature by measuring the gain ratio with respect to the class [13]. The calculation of gain ratio *GainR* is shown below, where *H* represents the entropy.

$$GainR(Class,Attribute)=(H(Class)-H(Class|Attribute))/H(Attribute)$$

The output of the algorithm is a list of descending features in terms of discriminative power. We call these features computational (bottom-up) features which can be added to the feature list for building the classification model.

As an example, we explored methods to extract and select computationally-derived linguistic features in order to improve the performance of deception detection in political speech [14]. We extracted linguistic features from different language tools and used feature selection techniques to select the optimal feature set. The selected features included both theoretically expected features (e.g., negative emotion tone) and empirically-derived features (e.g., narrative and cohesion). The results show that using computationally-derived features can significantly improve deception detection performance compared with a theoretical approach that uses a limited set of features.

C. HSCB dimensions layer

The SocialCube is designed to be adaptable and applicable to extracting information about many different kinds of HCSB dimensions (e.g., affect, deception, fatalism vs. mastery, group identity, etc.). In our previous research, we have shown that how linguistic features can be used to assess the deception dimension [14]. We provide more examples below on how linguistic features can be uniquely combined to assess a given HSCB dimension.

1) Affect

Perhaps one of the most important social and cultural dynamics for humans is their sense of emotion [15]. Emotion reflects not only how an individual is reacting to ongoing events, but can also reflect to how an individual generally views the world and his/her place in it. While emotion was long ignored by cognitive psychologists, a wide preponderance of data suggests that understanding an individual or group's emotional state can provide important insight and prediction into their decision-making, cognitive responses, and future behavior [16].

Although emotion is often assumed to be only communicated nonverbally [17], a number of recent studies suggest that humans convey their emotions in text-based communication, such as emails, blogs, instant messaging, and other forms of textual communication through linguistic cues. In one study [18], for example, individuals were asked to communicate only by text, and one partner was induced to feel sad before the interaction. Under these conditions, their partner was able to detect the negative emotion in the emotionally induced participant, indicating that emotion can be detected in text-based communication. Importantly for the present research, these data suggest that emotions can be detected from text-based communication.

There are specific linguistic patterns of emotional expression in verbal content, and there are a number of

established tools that can extract relevant emotional content, including the LIWC program and the Dictionary of Affect in Language program. Using these tools, Hancock and colleagues [19] have found that when people are sad they tend to use fewer words, disagree more, use more negative-affect words, and respond more slowly.

Taken together, these data suggest that emotional indicators or an individual or group can be extracted from verbal content present in text-based communication, and that these features, dynamically tracked over time, can predict emotionality of an individual or even a group.

2) Novel dimensions

As we note above, we are designing the system to be an extensible framework for identifying additional HSCB dimensions, even ones that cannot currently be conceptualized or predicted. Because we can combine the basic building blocks from our linguistic analysis to develop new HSCB dimensions, we believe the system is extremely powerful and adaptable.

For example, fatalism vs. mastery is an important phenomenon. People vary along the degree to which they feel that they have mastery over their life conditions versus a sense that they have little control over their lives. This sense of one's ability to control events around them has been measured by a number of different psychological and cultural dimensions, including learned helplessness [20] or fatalism. When individuals, groups or cultures feel that they have no control over their circumstances, they are said to be in a condition of learned helplessness, to have an external locus of control and a fatalistic set of beliefs, in which their fate is predetermined, based on luck and pessimism. Understanding an individual, group or culture's sense of fatalism can be an important indicator of that entity's overall psychological and cultural make-up, an understanding that can be important from an operational and intelligence point of view.

For another example, the group identity of a set of individuals may be important. Are they part of the same in-group, or are they enemies that make up two distinct out-groups? To analyze this question, the language of the individuals could be analyzed in terms of their pronoun use and whether first person plural (we) is used, signaling a common identity, or whether there is much more third person plural (they, them), signaling in-out group dynamics.

IV. TEXT CUBE

Data cube is a new way to organize data in multiple dimensions and multiple hierarchies for efficient information query and visualization from multiple perspectives [21]. A data cube allows data to be aggregated and viewed in multiple dimensions. It is defined by dimensions and measures (or facts). In general terms, dimensions are the perspectives with respect to which an organization wants to keep records (e.g., by time, by location, etc.). Each dimension may have a table associated with it called a dimension table. Measures are numerical measures that are quantities by which we want to analyze relationships between dimensions.

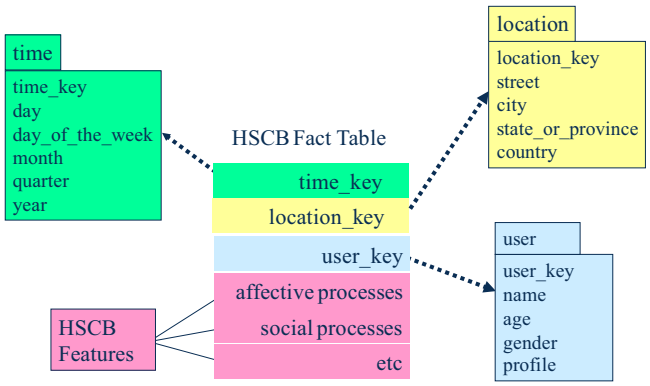

Fig. 3. Star schema of the data warehouse

A. Star schema

The star schema is a type of multidimensional model for the data cube. In a star schema, there are one or more fact tables referencing any number of dimension tables. The fact table contains the names of the facts (measures), as well as keys to each of the related dimension tables.

We have designed a star schema to store the extracted linguistic features for different HSCB dimensions. They are stored as measures in the Fact table. Fig. 3 shows a star schema design of our HSCB data warehouse. The fact table contains keys to dimensions such as time, location, and user. It also contains HSCB measures such as affective process and social processes calculated by LIWC. Note that some LIWC measures have hierarchical relationships. For example, "affective processes" can be divided to "positive emotion" and "negative emotion"; and "negative emotion" can be further divided into "anger", "anxiety", and "sadness".

B. Data Cube architecture

Based on star schema, we have designed a data cube architecture to allow users to conveniently view aggregated statistics of HSCB linguistic measures along different dimensions such as time and location, see Fig. 4. This type of data cube is also called "text cube".

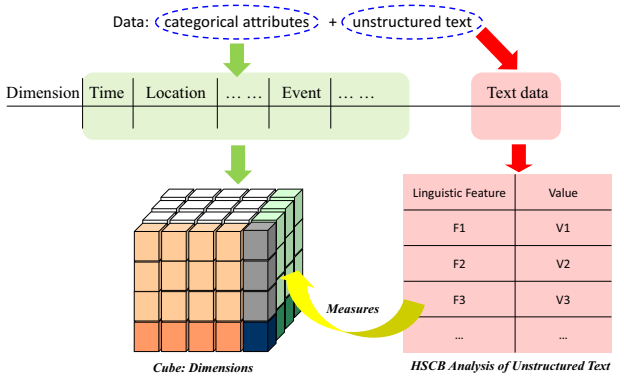

Fig. 4. Design of HSCB textual cube

We can specify the features that could be used as HSCB measures using a configuration file. For example, we can specify the features in the "affective processes" category of LIWC (e.g., positive emotion, negative emotion) and build an affect cube. Similarly, we can specify the features related to deception (e.g., first person singular, exclusive words, negative emotion) and build a deception analysis cube.

With the HSCB measures and dimensions defined we are now able to view the text cube and perform analysis. This includes slicing, dicing and drilling through cube cells. To demonstrate the text cube capability, we designed and implemented an interface to view the cubes (see Fig. 7 for an example).

V. DATA MINING APPROACH FOR HSCB MODELING

The text cube architecture also supports the development of prediction models using the data from the cells in a data cube. These models built on the summarized statistics represent higher level of knowledge representation. In the following, we introduce the data mining approach for HSCB modeling.

Data mining techniques allow us to select important linguistic features and to build prediction models for each HSCB dimension using the selected features. Data mining techniques, such as classification and clustering, will provide in-depth knowledge for each HSCB dimension and will complement the multidimensional data cube analysis.

We have designed a data mining solution for HSCB analysis by leveraging IAI's Agent-Based Data Miner (ABMiner) [22][23]. ABMiner supports full data mining cycle, including data set preparation, model discovery, and model deployment. For data set preparation, ABMiner provides a query designer which helps the user to retrieve data from various relational databases. For model discovery, ABMiner provides more than 400 machine learning algorithms (e.g., classification and clustering algorithms) aggregated from IAI's machine learning projects and open sources libraries such as Weka [24]. These algorithms allow users to build HSCB models (e.g., event detection models). ABMiner dynamically visualizes the model building process and the performance (e.g., accuracy) of each model. Users can compare the models and select the best model for deployment. Fig. 5 shows the screenshot of ABMiner.

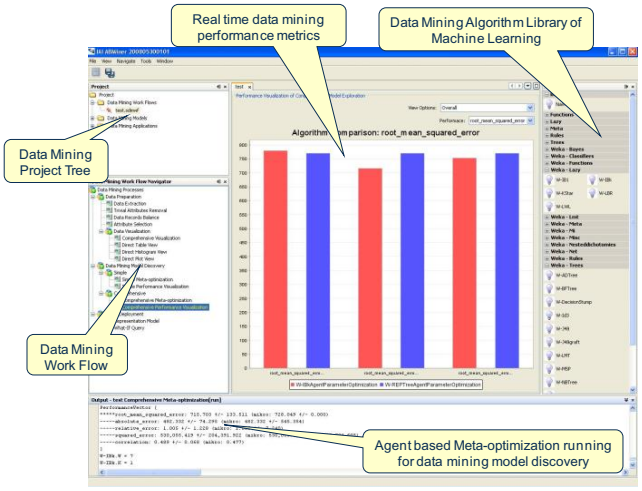

Fig. 5. ABMiner Screenshot

Here we focus on classification algorithms for HSCB modeling (e.g., building classifiers for event detection). Representative classification algorithms in ABMiner include:

- libSVM – is an efficient algorithm for support vector classification [25].
- IBK – is a K-nearest neighbours classifier. It can select appropriate value of K based on cross-validation. It can also do distance weighting [26].
- REPTree – is a fast decision tree learner. It is a mixture of decision tree and linear regression, where each leaf node corresponds to a linear regression algorithm [27].

ABMiner employs the *k*-fold cross validation method to evaluate classification models. In *k*-fold cross validation, the data set is divided into *k* subsets. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times. The variance of the resulting estimate is reduced as *k* is increased.

VI. CASE STUDIES

A. Data

We developed a data collection system to extract live tweets and filter them by topics. Using the system, we extracted ~64,000 tweets on the Egyptian revolution during January 25, 2011 and February 11, 2011. This data gives us an example to study the linguistic features of political events and their potential predictive power of events.

B. HSCB linguistic analysis

To assist theoretical and computational analysis of different HSCB dimensions, we implemented a HSCB feature extraction tool. This tool has two major components: the language translation agent and the linguistic feature extraction agent. The language translation agent is responsible for detecting and translating foreign languages into English. Then the linguistic feature extraction agent processes the translated tweets using natural language processing tools and extract HSCB features.

Using the HSCB feature extraction tool, we conducted case studies on affect analysis of political events (e.g., Libya civil war [28]). Here we discuss the HSCB text analysis using the tweets on the Egyptian revolution. We extracted linguistic features on affect (i.e., negative and positive emotions) using the LIWC method, and plotted the linguistic features as a function of key events during the revolution.

We found that affect is highly predictive of major events and reflective of moods in the Egyptian populace. As shown in Fig. 6, negative emotion levels in the tweets were the highest on Feb 2, which corresponds to hundreds of casualties that occurred that day. Also, the positive emotion was the highest on Feb 4, which corresponds to the "Day of Departure". These results provide some initial evidence for the utility of HSCB linguistic analysis for event reporting, and would allow intelligence assessment of cultural dynamics without having to put resources into the field.

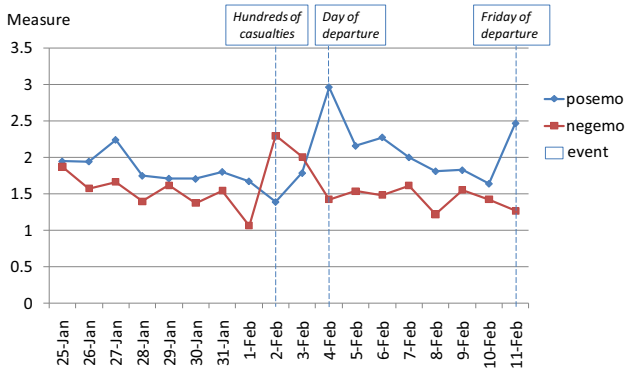

Fig. 6. Linguistic analysis of tweets on Egypt's revolution

These initial analyses are very promising, and they represent only a fraction of the kinds of inferences that can be gleaned from unstructured texts, such as tweet. While the emotional representations in Fig. 6 reflect current emotion in the population in response to specific events, other dimensions of language could be more predictive of events likely to take place. One dimension of interest that we will analyze for our subsequent report is swearing behavior, which we expect to be a measure of frustration and anger that can predict when large-scale protests will take place. This kind of information will be extremely valuable for intelligence assessment in closed societies where traditional polling or the use of other intelligence assets for assessing HSCB is severely limited.

C. *Text cube multidimensional analysis*

To support multidimensional analysis, we designed and implemented one text cube looking at tweets on the topic of the Egypt revolt. The dimensions include the location and time. The HSCB measures used are the LIWC features extracted from the tweets.

Fig. 7 shows the Affective Processes cube for the Egypt revolt. The horizontal dimension is the cities from which the tweets originated from and the vertical dimension is the time period. Its measures include LIWC features such as Affective Processes, Positive Processes, Negative Processes, Anger, Anxiety, Sadness, Religion, and Social.

	Time					
	All Periods					
	Measures					
Location	Affective Processes	Positive Emotion	Negative Emotion	Anger	Anxiety	Sadness
All Locations	3.491	1.96	1.521	0.775	0.202	0.308
	3.634	2.018	1.491	0.825	0.188	0.131
"Arlington, VA"	0.893	0	0.893	0.893	0	0
"Baltimore, MD"	4.35	0	4.35	0	4.35	0
"Boston, MA"	4.559	3.633	0.926	0	0	0
"Cairo, Egypt / Dubai"	5.904	3.96	1.908	1.096	0.195	0.268
"Cairo, Egypt"	5.156	2.799	2.358	1.896	0	0.463
"Cambridge, UK"	5.968	2.38	3.588	1.138	1.138	0
"Doha, Qatar"	1.779	1.118	0.727	0.447	0	0.28
"Edinburgh, UK"	2.6	1.15	1.45	0	0	1.45
"Edmonton, Alberta, Canada"	0	0	0	0	0	0
"Exeter, UK"	0	0	0	0	0	0
"Flemington, New Jersey"	0	0	0	0	0	0
"Fucknutsville, DC"	5.88	0	5.88	5.88	0	0
"Goes, The Netherlands"	3.992	2.056	1.936	0.877	0	0.458
"Hell's Breakfast Nook, NYC"	4.331	2.345	1.394	0.558	0.309	0.263
"Jakarta, Indonesia"	0	0	0	0	0	0

Fig. 7. Affective Processes Cube for the Egypt Revolt

	Location						
		"Arlington, VA"		"Baltimore, MD"		"Bo	
	Measures	Measures		Measures		Mea	
Time	Positive Emotion	Negative Emotion	Positive Emotion	Negative Emotion	Positive Emotion	Negative Emotion	Po
2011-01-25	0	0					
2011-01-26	0	3.845					
2011-01-27							
2011-01-28	1.667	0.391	0	3.125			
2011-01-29	0	0			0	4.35	
2011-01-30			0	0			
2011-01-31			0	0			
2011-02-01	5.663	0					
2011-02-02	0	0					
2011-02-03	2.894	1.428	0	0			
2011-02-04	0	1.667					
2011-02-05	2.563	0					
2011-02-06	3.253	1.853					
2011-02-07	0	5.553	0	0			
2011-02-08	4.228	2.792					
2011-02-09	0	0					
2011-02-10	1.283	2.768					
2011-02-11	2.675	1.083	0	0			

Fig. 8. Text cube showing measures by location and time

We can expand the time dimension to drill down to specific days. We can also change the text cube view by switching the horizontal and vertical dimensions. Fig. 8 shows another view of the text cube where time is the horizontal dimension and location is the vertical dimension.

Again, we can aggregate the measures for all locations by "shrinking" the location dimension. This will generate the dataset that is suitable for time series trend analysis of measures (e.g., positive emotion, negative emotion, etc.) regardless of specific locations. Our text cube interface also provides plotting capability to plot aggregated measures for all locations on each day (e.g., a plot similar to Fig. 6).

D. *SocialCube for Predictive Modeling: Predicting Violence in the Egyptian Revolution*

We studied the feasibility of data mining algorithms for predictive modeling. We used event detection in texts as an example modeling task. Event detection can be treated as a classification problem, where a model or classifier is constructed to predict the categorical labels of events (e.g., "large-scale" vs. "small-scale", or "violent" vs. "non-violent").

In a classification problem for event detection, the inputs are the language features (e.g., LIWC features) stored in the data cube and the output is the categorical label. To associate the inputs with the output, models need to be developed using classification algorithms such as support vector machines and neural networks. Then a model with the best performance (e.g., accuracy) is selected for deployment. As new texts come in, the model will apply the language features of the texts as inputs and predict the output labels.

We designed the data mining problem of predicting the scale and degree of violence in the Egyptian Revolution. Given the LIWC features measured from the tweets collected on each day (e.g., the positive and negative emotions shown in Fig. 6), our goal was to predict whether there were large-scale and violent events for each day. To obtain the ground truth, we referred to the Timeline of the 2011 Egyptian Revolution [29]. Based on the descriptions of protests and conflict events during January 25, 2011 and February 11, 2011, we manually coded the categorical labels of events for each day. TABLE I. shows the input LIWC features together with two types of prediction labels: scale and degree of violence.

TABLE I. SAMPLE DATA MINING DATASET

Date	1st person singular	Social	Affect	Positive emotion	Negative emotion	Anger	Exclusive words	Motion	Religion	Label: scale	Label: violence
25-Jan	0.96	4.32	3.81	1.94	1.87	1.08	0.92	1.73	0.58	large-scale	non-violent
26-Jan	0.82	4.80	3.50	1.94	1.57	0.85	0.66	1.30	1.00	large-scale	violent
27-Jan	0.57	5.06	3.90	2.24	1.66	1.10	0.96	1.61	1.29	small-scale	non-violent
28-Jan	0.75	4.18	3.18	1.75	1.40	0.92	0.82	1.61	0.58	large-scale	non-violent
29-Jan	0.69	5.12	3.35	1.71	1.61	0.58	0.81	1.31	0.50	large-scale	non-violent
30-Jan	1.19	5.48	3.10	1.70	1.37	0.58	0.95	1.85	0.42	small-scale	non-violent
31-Jan	1.24	6.29	3.38	1.80	1.54	0.71	0.88	1.81	0.48	large-scale	non-violent
1-Feb	0.66	4.91	2.71	1.67	1.06	0.62	1.01	1.61	0.41	small-scale	violent
2-Feb	1.10	6.18	3.65	1.39	2.29	1.22	1.32	1.85	0.73	large-scale	violent
3-Feb	1.24	6.38	3.85	1.78	2.00	1.22	1.13	1.89	0.59	large-scale	violent
4-Feb	1.00	5.43	4.37	2.96	1.42	0.64	1.01	1.98	0.57	large-scale	non-violent
5-Feb	0.72	4.37	3.70	2.16	1.53	0.66	0.68	1.26	1.05	large-scale	non-violent
6-Feb	0.88	5.28	3.78	2.27	1.48	0.85	0.95	1.56	0.48	large-scale	non-violent
7-Feb	0.97	4.98	3.61	2.00	1.61	0.70	1.26	0.86	0.76	large-scale	non-violent
8-Feb	0.70	5.12	3.01	1.81	1.22	0.70	0.70	1.88	1.02	large-scale	non-violent
9-Feb	0.76	4.10	3.43	1.82	1.55	0.74	0.67	1.11	1.02	large-scale	violent
10-Feb	0.69	5.13	3.06	1.64	1.42	0.70	0.92	1.78	0.46	large-scale	non-violent
11-Feb	1.27	6.23	3.72	2.46	1.26	0.46	0.75	2.01	0.38	large-scale	non-violent

We used the data in TABLE I. to train different classifiers, including libSVM, REPTree, and IBK. We tested the performance of these classifiers using 10-fold cross-validation. TABLE II. shows the cross validation results measured by classification accuracy. Overall, these classifiers are able to detect the scale and degree of violence with reasonable accuracy, despite the small number of training examples. The accuracy is higher than baseline/chance accuracy for scale and violence (50%). This is perhaps not surprising as the training set contains multiple linguistic and psychological features (e.g., positive and negative emotions) that are predictive of events, as already shown in Section VI.B.

TABLE II. CROSS-VALIDATION RESULTS MEASURED BY ACCURACY

Categorical Label	**libSVM**	**REPTree**	**IBK**
Large-scale vs. small-scale	83.33%	83.33%	73.33%
Violent vs. non-violent	73.33%	73.33%	60.00%

The data mining example shown here is just the tip of the iceberg of possible studies that can be conducted for modeling cyber activism. The rapid growth of social media and unstructured data on the Internet has created unprecedented opportunities for understanding and predicting social dynamics, and the reporting of news has made it possible to track in near real-time events taking place around the globe. Integrating social data emerging from the Internet with real-world events data has the potential to predict and mitigate future conflicts.

VII. CONCLUSIONS

We have introduced the concept of SocialCube for analyzing cyber behaviors or human social cultural behaviors (HSCB) in unstructured text. We have demonstrated the feasibility of SocialCube for HSCB data collection and analysis. Our key contributions include:

- **HSCB feature analysis**. We have developed a comprehensive HSCB linguistic feature analysis framework that allows for an extensible set of HSCB dimensions that can be developed on an as needed basis. The framework provides generic linguistic features from LIWC, Stanford POS tagger, and Coh-Metrix. Together, these tools produce the basic elements for HSCB dimensions such as affect, deception, and sense of fatalism, and for any future HSCB dimension.
- **Data cube architecture for multidimensional analysis**. We have developed a data cube architecture to summarize and navigate structured data (dimensions) together with unstructured text data (measures) for efficient query and analysis. The data cube has an underlying star schema to store the high dimensional HSCB data. The star schema has a fact table which contains the linguistic measures as well as keys to each of the related dimension tables such as location and time. With the schema defined, users are able to view the cube models and perform analysis.
- **Data mining for predictive HSCB modeling**. SocialCube leverages IAI's ABMiner data mining platform for HSCB modeling. ABMiner integrates hundreds of data mining algorithms (e.g., clustering, classification, anomaly detection) from IAI's machine learning projects and open sources libraries. These algorithms provide the capability to build prediction models using linguistic features. We used political event detection in texts as an example modeling task. The results show that accurate prediction models (e.g., accuracy over 80%) could be built using HSCB linguistic features.

Using social language processing and data cube for HSCB analysis is an emerging area for research. Many problems in this field remain to be studied. In light of this research, some topics that need further investigation include 1) design data collection methods for in-situ collection of large-scale social media data for cyber behaviors analysis, 2) expand the HSCB analysis framework to include emerging new HSCB dimensions, 3) incorporate more functions on text cube (e.g.,

topic modeling [30], online analysis [31], keyword based exploration [32]), 4) develop advanced prediction methods for predicting future behaviors and events, and 5) enhance user interface and visualization methods.

As social media continues to grow, we will also address data integration and scalability issues. Nowadays, cloud computing technologies, such as Hadoop/MapReduce, Pig, Hive, are well-known for processing big data [33]. We will design strategies to distribute computationally expensive tasks (e.g., data collection, stream cubing, mining, etc.) over a cloud system. The goal is to develop an efficient and flexible system for processing large-scale data streams.

ACKNOWLEDGMENTS

We thank the reviewers for the valuable comments. Part of this research was funded by Navy Grant # N00014-11-M-0102 awarded to IAI. An earlier version of this paper was presented at the Inter-Asia Roundtable on Cyberactivism, organized by Asia Research Institute at the National University of Singapore, in Singapore on August 30-31, 2012.

REFERENCES

[1] S.K. Numrich, and A. Tolk. Challenges for Human, Social, Cultural, and Behavioral Modeling. SCS M&S Magazine 1(1), January 2010.

[2] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh, "Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab, and Sub Totals", Data Mining and Knowledge Discovery, 1(1), pp. 29-53, 1997.

[3] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In International AAAI Conference on Weblogs and Social Media, Washington, D.C., 2010.

[4] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy, Fourth International AAAI Conference on Weblogs and Social Media, May 23-26, 2010, Washington, DC.

[5] K. Lerman, R. Ghosh, Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks, Fourth International AAAI Conference on Weblogs and Social Media, May 23-26, 2010, Washington, DC.

[6] J.W. Pennebaker, R.J. Booth, and M.E. Francis, (2007). Linguistic Inquiry and Word Count: LIWC. Austin, TX: LIWC www.liwc.net.

[7] K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

[8] A.C. Graesser, D.S. McNamara, M.M. Louwerse, and Z. Cai. Coh-Metrix: Analysis of text on cohesion and language. Behavioral Research Methods, Instruments, and Computers, 36, 193–202, 2004.

[9] M. L. Newman, J.W. Pennebaker, D.S. Berry, J. M. Richards. Lying words: Predicting deception from Linguistic styles. Personality and Social Psychology Bulletin, 29, 665-675, 2003.

[10] J.T. Hancock, L. Curry, S. Goorha, and M.T. Woodworth. On lying and being lied to: A linguistic analysis of deception. Discourse Processes, 45, 1-23, 2008.

[11] G.D. Bond, & A. Y. Lee. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. Applied Cognitive Psychology, 19, 313-329, 2005.

[12] P.S. Keila and D.B. Skillicorn. Detecting unusual and deceptive communication in email. IBM Centers for Advanced Studies Conference, 2005.

[13] J. Quinlan. C4.5 Programs for Machine Learning. Morgan Kauman Press, 1993.

[14] X. Liu, J. Hancock, G. Zhang, R. Xu, D. Markowitz, and N. Bazarova. Exploring linguistic features for deception detection in unstructured text. Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, Hawaii International Conference on System Sciences (HICSS), January 4-7, 2012.

[15] P. Ekman. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. New York: Norton & Company Inc., 2001.

[16] J.A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161-1178, 1980.

[17] A. Mehrabian. Nonverbal communication. Chicago: Aldine-Atherton, 1972.

[18] J.T. Hancock, C. Landrigan, and C. Silver. Expressing emotion in text. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2007), 929 – 932.

[19] J.T. Hancock, K. Gee, K. Ciaciaco, and J. Mae. I'm sad you're sad: Emotional contagion in CMC. Proceedings of the ACM conference on Computer-Supported Cooperative Work (CSCW 2008).

[20] M.E.P. Seligman. Helplessness: On Depression, Development, and Death. San Francisco: W.H. Freeman, 1975.

[21] C.X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao, Text Cube: Computing IR Measures for Multidimensional Text Database Analysis, Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.

[22] X. Liu, K. Tang, J.R. Buhrman, and H. Cheng. An Agent-based Framework for Collaborative Data Mining Optimization. Proceedings of the IEEE International Symposium on Collaborative Technologies and Systems (CTS), Chicago, Illinois, USA, May 17 - 21, 2010.

[23] K. Tang, X. Liu, Y. Tang, V. Manikonda, J. Buhrman, and H. Cheng. ABMiner: A Scalable Data Mining Framework to Support Human Performance Analysis, in V. Duffy (eds.), Advances in Applied Digital Human Modeling, CRC Press, 2010.

[24] http://www.cs.waikato.ac.nz/ml/weka/

[25] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using the second order information for training SVM. Journal of Machine Learning Research 6, 1889-1918, 2005.

[26] D. Aha, D. Kibler. Instance-based learning algorithms. Machine Learning. 6:37-66, 1991.

[27] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, San Francisco, 2 edition, 2005.

[28] C. Brown, J. Frazee, D. Beaver, X. Liu, F. Hoyt, J. Hancock. Evolution of Sentiment in the Libyan Revolution. White Paper at https://webspace.utexas.edu/dib97/libya-report-10-30-11.pdf. Oct. 30, 2011.

[29] http://en.wikipedia.org/wiki/Timeline_of_the_2011-2012_Egyptian_revolution

[30] D. Zhang, C. Zhai and J. Han, Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases, Proc. 2009 SIAM Int. Conf. on Data Mining (SDM'09), Sparks, NV, April 2009.

[31] D. Zhang, C. Zhai and J. Han, MiTexCube: MicroTextCluster Cube for Online Analysis of Text Cells, Proc. 2011 NASA Conf. on Intelligent Data Understanding (CIDU'11), Mountain View, CA, Oct. 2011.

[32] B. Zhao, C.X. Lin, B. Ding, J. Han, TEXplorer: Keyword based Object Ranking and Exploration in Multidimensional Text Databases, Proc. 2011 Int. Conf. on Information and Knowledge Management (CIKM'11), Glasgow, UK, Oct. 2011.

[33] J. Lin, D. Ryaboy, and K. Weil. Full-text indexing for optimizing selection operations in large-scale data analytics. MAPREDUCE Workshop, 2011.