COGNITIVE BIG DATA: SURVEY AND REVIEW ON BIG DATA ANALYTICS FOR ONLINE SOCIAL MEDIA

S. Gandhimathi,
Research Scholar,
PG & Research Department of Computer Science,
Govt. Arts College (Autonomous),
Karur,Tamilnadu,India.

Dr. V. Baby Deepa,
Assistant Professor,
PG & Research Department of Computer Science,
Govt. Arts College (Autonomous),
Karur,Tamilnadu,India.

Abstract: Big data and the social media is the emerging technology now days. The social media has become main communication way to connect with our family, friend and colleagues. People are sharing all information in many forms such as text, audio as well as video on the social networking site to shares their feeling. But it requires a more thorough discussion beyond the very common 3V (velocity, volume, and variety) approach. We established an expert group to re-discuss the notion of Big Data, identify new characteristics, and re-think what actually really is new in the idea of Big Data by analyzing over 100 literature resources. We identified typical baseline scenarios (traffic, business processes, retail, health, and social media) as starting point, from which we explored the notion of Big Data from a different viewpoint. We introduce a fully new way of thinking about Big Data, and coin it as the trend of 'Cognitive Big Data'. The publications introduce a basic framework for our research results. However, this work remains work-in-progress, and we will continue with a refinement of the Cognitive Big Data Framework in one future publication.

Keywords: Big Data, Scenarios, Data Research, Cognitive Big Data, Theory, Big Social Data, Social Big Data, Social computing, Classification Big Social Data analysis

I.INTRODUCTION

This paper will argue that this 'nothing new' way of thinking about Big Data research has a serious defect, because there are in fact several issues raised by this research, only some of them new. This paper documents the fundamental concepts with reference to big data. The fast rise of big data has left several unprepared. Within the past, new technological developments initial appeared in technical and educational publications. The information and synthesis later seeped into alternative avenues of data mobilization, together with books. The quick evolution of big data technologies and also the prepared acceptance of the idea by public and personal sectors left very little time for the discourse to develop and mature within the educational domain. Authors and practitioners leapfrogged to books and alternative electronic media for prompt and wide course of their work on huge information. Thus, one finds many books on big data, together with big data for Dummies, however not enough elementary discourse in educational publications. The leapfrogging of the discourse on big data to additional widespread retailers implies that a coherent understanding of the idea and its word is nevertheless to develop. As an example, there's very little agreement round the elementary question of however big the data has got to be to qualify as 'big data'.

Thus, there exists the necessity to document within the educational press the evolution of big data ideas and technologies. A key contribution of this paper is to originate the oft-neglected dimensions of big data. It ignores the most important part of big data, which is unstructured and is on the market as audio, images, video, and unstructured text. It's calculable that the analytics ready structured information forms solely low set of big data. The unstructured data, particularly data in video format, is that the largest part of big data that's solely part archived. This paper is organized as follows. We start the paper by shaping big data. We tend to highlight the actual fact that size is barely one among many dimensions of big data.

Alternative characteristics, like the frequency with that information are generated, are equally vital in shaping big data. We tend to then expand the discussion on numerous sorts of big data, particularly text, audio, video, and social media. We tend to apply the analytics lens to the discussion on big data. The discussion has remained centered on correlation, ignoring the additional nuanced and concerned discussion on exploit. We tend to conclude by highlighting the expected developments to comprehend within the close to future in big data analytics.

II. BIG DATA IN THE CURRENT ENVIRONMENTS OF ENTERPRISE AND TECHNOLOGY

In 2012, 2.5 quintillion bytes of data were generated daily, and 90% of current data worldwide originated in the past two years ([20] and Big Data, 2013). During 2012, 2.2 million TB of new data are generated each day. In 2010, the market for BigData was $3.2 billion, and this value is expected to increase to $16.9 billion in 2015 [20]. As of July 9, 2012, the amount of digital data in the world was 2.7 ZB [11]; Facebook alone stores, accesses, and analyzes 30 + PB of user-generated data [16]. In 2008, Google was processing 20,000 TB of data daily [44]. To enhance advertising, Akamai processes and analyzes 75 million events per day [45]. Walmart processes over 1 million customer transactions, thus generating data in excess of 2.5 PB as an estimate.

More than 5 billion people worldwide call, text, tweet, and browse on mobile devices [46]. The amount of e-mail accounts created worldwide is expected to increase from 3.3 billion in 2012 to over 4.3 billion by late 2016 at an average annual rate of 6% over the next four years. In 2012, a total of 89 billion e-mails were sent and received daily, and this value is expected to increase at an average annual rate of 13% over the next four years to exceed 143 billion by the end of 2016 [47]. In 2012, 730 million users (34% of all e-mail users) were e-mailing through mobile devices. Boston.com[47] reported that in 2013, approximately 507 billion e-mailswere sent daily.

Currently, an e-mail is sent every 3.5 × 10−7 seconds. Thus, the volume of data increases per second as a result of rapid data generation. Growth rates can be observed based on the daily increase in data. Until the early 1990s, annual growth rate was constant at roughly 40%. After this period, however, the increase was sharp and peaked at 88% in 1998 [7]. Technological progress has since slowed down. In late 2011, 1.8 ZB of data were created as of that year, according to IDC [21]. In 2012, this value increased to 2.8 ZB. Globally, approximately 1.2 ZB(1021) of electronic data are generated per year by various sources [7]. By 2020, enterprise data is expected to total 40 ZB, as per IDC [12]. Based on this estimation, business to-consumer (B2C) and internet-business-to-business (B2B) transactions will amount to 450 billion per day. Thus, efficient management tools and techniques are required.

III. BIG DATA MANAGEMENT

The architecture of Big Data must be synchronized with the support infrastructure of the organization. To date, all of the data used by organizations are stagnant. Data is increasingly sourced from various fields that are disorganized and messy, such as information from machines or sensors and large sources of public and private data. Previously, most companies were unable to either capture or store these data, and available tools could not manage the data within a reasonable amount of time. However, the new Big Data technology improves performance, facilitates innovation in the products and services of business models, and provides decision making support [8, 48]. The Big Data technology aims to minimize Hardware and processing costs and to verify the value of Big Data before committing significant company resources.

Properly managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can be applied In various complex scientific disciplines (either single or interdisciplinary), including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry. In the following section, we briefly discuss data management tools and propose a new data life cycle that uses the technologies and terminologies of Big Data.

3.1. Management Tools. With the evolution of computing technology, immense volumes can be managed without requiring supercomputers and high cost. Many tools and techniques are available for data management, including Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort [3]. However, companies must develop special tools and technologies that can store, access, and analyze large Amounts of data in near-real time because Big Data differs from the traditional data and cannot be stored in a single machine. Furthermore, Big Data lacks the structure of traditional data.

For Big Data, some of the most commonly used tools and techniques are Hadoop, MapReduce, and Big Table. These innovations have redefined data management because they effectively process large amounts of data efficiently, cost effectively, and in a timely manner. The following section describes Hadoop and MapReduce in further detail, as well as the various projects/frameworks that are related to and suitable for the management and analysis of Big Data.

3.2. Hadoop: Hadoop [49] is written in Java and is a top-level Apache project that started in 2006. It emphasizes discovery from the perspective of scalability and analysis to realize near-impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system. Presently, it is used in large amounts

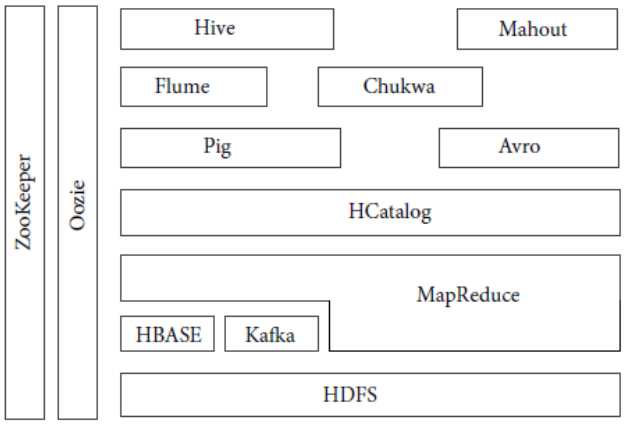

Figure 3: Hadoop ecosystem.

of data. With Hadoop, enterprises can harness data that were previously difficult to manage and analyze. Hadoop is used by
approximately 63% of organizations to manage huge number of unstructured logs and events (Sys. con Media, 2011). In particular, Hadoop can process extremely large volumes of data with varying structures (or no structure at all). The following section details various Hadoop projects and their links according to [12, 50–55].

Hadoop is composed of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka; however, the most common components and well-known paradigms are Hadoop Distributed File System (HDFS) and MapReduce for Big Data. Figure 3 illustrates the Hadoop ecosystem, as well as the relation of various components to one another.

HDFS. This paradigm is applied when the amount of data is too much for a single machine. HDFS is more complex than other file systems given the complexities and uncertainties of networks. The cluster contains two types of nodes. The first node is a name-node that acts as a master node. The second

node type is a data node that acts as slave node. This type of node comes in multiples. Aside from these two types of nodes, HDFS can also have secondary name-node. HDFS stores files in blocks, the default block size of which is 64MB. All HDFS files are replicated in multiples to facilitate the parallel processing of large amounts of data.

HBase: HBase is a management system that is open-source, versioned, and distributed based on the BigTable of Google. This system is column- rather than row-based, which accelerates the performance of operations over similar values across large data sets. For example, read and write operations involve all rows but only a small subset of all columns. Hbase is accessible through application programming interfaces (APIs) such as the Thrift, Java, and representational state transfer (REST). These APIs do not have their own query or scripting languages. By default, HBase depends completely on a ZooKeeper instance.

ZooKeeper. ZooKeeper maintains, configures, and names large amounts of data. It also provides distributed synchronization and group services.This instance enables distributed processes to manage and contribute to one another through a name space of data registers (*z*-nodes) that is shared and hierarchical, such as a file system. Alone, ZooKeeper is a
distributed service that contains *master* and *slave* nodes and stores configuration information.

HCatalog. HCatalog manages HDFS. It stores metadata and generates tables for large amounts of data. HCatalog depends on Hive metastore and integrates it with other services, including MapReduce and Pig, using a common data model. With this data model, HCatalog can also expand to HBase. HCatalog simplifies user communication using HDFS data and is a source of data sharing between tools and execution platforms.

Hive. Hive structures warehouses in HDFS and other input sources, such as Amazon S3. Hive is a subplatform in the Hadoop ecosystem and produces its own query language (HiveQL). This language is compiled by MapReduce and enables user-defined functions (UDFs). The Hive platform is primarily based on three related data structures: tables, partitions, and buckets. Tables correspond to HDFS directories
and can be distributed in various partitions and, eventually, buckets.

Pig. The Pig framework generates a high-level scripting language (Pig Latin) and operates a run-time platform that enables users to execute MapReduce on Hadoop. The Pig is more elastic than Hive with respect to potential data format given its data model. Pig has its own data type, *map*, which represents semi structured data, including JSON and XML.

Mahout. Mahout is a library for machine-learning and data mining. It is divided into four main groups: collective filtering, categorization, clustering, and mining of parallel frequent patterns. The Mahout library belongs to the subset that can be executed in a distributed mode and can be executed by MapReduce.

Oozie. In the Hadoop system, Oozie coordinates, executes, and manages job flow. It is incorporated into other Apache Hadoop frameworks, such as Hive, Pig, Java MapReduce, Streaming MapReduce, and Distcp Sqoop. Oozie combines actions and arranges Hadoop tasks using a directed acyclic graph (DAG). This model is commonly used for various tasks.
Avro. Avro serializes data, conducts remote procedure calls, and passes data from one program or language to another. In this framework, data are self-describing and are always stored based on their own schema because these qualities are particularly suited to scripting languages such as Pig.

Chukwa. Currently, Chukwa is a framework for data collection and analysis that is related to MapReduce and HDFS. This framework is currently progressing from its development stage. Chukwa collects and processes data from distributed systems and stores them in Hadoop. As an independent module, Chukwa is included in the distribution of Apache Hadoop.

IV. BIG DATA ANALYTICS

Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics are a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways [12,13, 15, 16]. The main goal of the big data analytic is to help organization to make better business decision, future prediction, analysis large numbers of transactions that done in organization and update the form of data that organization is used. Example of big data Analytics are big online business website like Flipkart, snapdeal uses Facebook or Gmail data to view the customer information or behaviour. Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independently or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions. It helps us to uncover hidden patterns, unknown correlations, market trends, customer preferences etc. It leads us to more effective marketing, revenue opportunities, better customer service etc.

Big Data can be analyzed through predictive analytics, text analytics, statistical analytics and data mining [1,2,4]. Types of big data analytics are: Prescriptive: - This type of analytics helps to decide what actions should be taken. It's very valuable, but not used largely. It focuses on answer specific questions like, hospital management, diagnosis of cancer patients, diabetes patients that determine where to focus treatment. Predictive: - This type of analytics helps to predict the future or what might be happen. For example, some companies use predictive analytics to take decision for sales, marketing, production, etc.

Diagnostic: - In this type look at past and analyze the situation, what happen in the past and why it happen. And how we can overcome this situation. For example weather prediction, customer behavioral analysis etc. Descriptive:-It describes what is happening currently and prediction near future. For example market analysis, compactions behavioral analysis etc. By using appropriate analytics organization can increase sales, increase customer service, and can improve operations. Predictive Analytics allow organizations to make better and faster decisions [1, 2, 4, 10].

4.1. Predictive Analytics

Predictive Analytics is a method through which we can extract information from existing data sets to predict future outcomes and trends and also determine patterns. It does not tell us what will happen in future. It forecasts what might happen in future with an acceptable level of reliability. It also includes what if-then-else scenarios and risk assessment. Applications areas of Predictive Analytics are [1, 2, 4, 10]:

CRM (Customer Relationship Management): Predictive analytics is useful in CRM in fields such as marketing campaigns, sales, customer services etc. The focus is to put their efforts effectively on analyzing a product in demand and predict customer's buying habits. Clinical Decision Support: Predictive Analytics help us to determine that patients which are at risk of developing certain conditions like diabetes, asthma, lifetime illness etc.

Collection Analytics: Predictive Analytics help financial institutions for the allocation in collecting resources by identifying most effective collection agencies, contact strategies etc. to each customer.

Cross Sell: An Organization that offers multiple products, Predictive Analytics can help to analyze customer's spending, their behavior etc. This can help to lead cross sales, that means selling additional products to current customers.

Customer Retention: As the number of competing services is increasing, businesses should continuously focus on maintaining customer satisfaction, rewarding loyal customers and minimize customer reduction. If Predictive Analytics are properly applied, it can lead to an active retention strategy by frequently examining customer's usage, spending and behavior patterns.

Direct marketing: When marketing consumer products and services, there is the challenge of keeping up with competing products and consumer behavior. Apart from identifying prospects, predictive analytics can also help to identify the most effective combination of product versions, marketing material, communication channels and timing that should be used to target a given consumer.

Fraud detection: Fraud is a big problem for many businesses and can be of various types: inaccurate credit applications, fraudulent transactions (both offline and online), identity thefts and false insurance. These problems plague firms of all sizes in many industries. Some examples of likely victims are credit card issuers, insurance companies, retail merchants, manufacturers, business-to-business suppliers and even service providers. Predictive analysis can help to identify high-risk, fraud candidates in business or the public sector.

Portfolio, product or economy-level prediction: These types of problems can be addressed by predictive analytics using time series techniques. They can also be addressed via machine learning approaches which transform the original time series into a feature vector space, where the learning algorithm finds patterns that have predictive power.

Risk management: When employing risk management techniques, the results are always to predict and benefit from a future scenario. Predictive analysis helps organizations or business enterprises to identify future risk, Natural Disaster and its effect. Risk management helps them to take correct decision on correct time.

Underwriting: Many businesses have to account for risk exposure due to their different services and determine the cost needed to cover the risk. For example, auto insurance providers need to accurately determine the amount of premium to charge to cover each automobile and driver. For a health insurance provider, predictive analytics can analyze a few years of past medical claims data, as well as lab, pharmacy and other records where available, to predict how expensive an enrollee is likely to be in the future. Predictive analytics can help underwrite these quantities by predicting the chances of illness, default, bankruptcy, etc. Predictive analytics can streamline the process of customer acquisition by predicting the future risk behaviour of a customer using application level data.

Social media research and applications:

Social media data is clearly the largest, richest and most dynamic evidence base of human behavior, bringing new opportunities to understand individuals, groups and society. Innovative scientists and industry professionals are increasingly finding novel ways of automatically collecting, combining and analyzing this wealth of data. Naturally, doing justice to these pioneering social media applications in a few paragraphs is challenging. Three illustrative areas are: business, bioscience and social science.

The early business adopters of social media analysis were typically companies in retail and finance. Retail companies use social media to harness their brand awareness, product/customer service improvement, advertising/ marketing strategies, network structure analysis, news propagation and even fraud detection. In finance, social media is used for measuring market sentiment and news data are used for trading. As an illustration, Bollen et al. (2011) measured sentiment of a random sample of Twitter data, finding that Dow Jones Industrial Average (DJIA) prices are correlated with the Twitter sentiment 2–3 days earlier with 87.6 percent accuracy. Wolfram (2010) used Twitter data to train a Support Vector Regression (SVR) model to predict prices of individual NASDAQ stocks, finding 'significant advantage' for forecasting prices 15 min in the future.

In the biosciences, social media is being used to collect data on large cohorts for behavioral change initiatives and impact monitoring, such as tackling smoking and obesity or

monitoring diseases. An example is Penn State University biologists (Salathe´ et al. 2012) who have developed innovative systems and techniques to track the spread of infectious diseases, with the help of news Web sites, blogs and social media.

Computational social science applications include: monitoring public responses to announcements, speeches and events, especially political comments and initiatives; insights into community behavior; social media polling of (hard to contact) groups; early detection of emerging events, as with Twitter. For example, Lermanetal. (2008) use computational linguistics to automatically predict the impact of news on the public perception of political candidates. Yessenov and Misailovic (2009) use movie review comments to study the effectiveness of various approaches in extracting text features on the accuracy of four machine learning methods—Naive Bayes, Decision Trees, Maximum Entropy and K-Means clustering. Lastly, Karabulut (2013) found that Facebook's Gross National Happiness (GNH) exhibits peaks and troughs in-line with major public events in the USA.

V. SOCIAL MEDIA OVERVIEW

For this paper, we group social media tools into:

• **Social media data**—social media data types (e.g., Social networking media, wikis, blogs, RSS feeds and news, etc.) and formats (e.g., XML and JSON). This includes data sets and increasingly important real-time data feeds, such as financial data, customer transaction data, telecoms and spatial data.

- **Social media programmatic access**—data services and tools for searching and scraping (textual) data from social networking media, wikis, RSS feeds, news, etc. These can be usefully subdivided into:
    - **Data sources, services and tools**—where data is accessed by tools which protect the raw data or provide simple analytics. Examples include: Google Trends, Social Mention, Social Pointer and Social-Seek, which provide a stream of information that aggregates various social media feeds.
    - **Data feeds via APIs**—where data sets and feeds are accessible via programmable HTTP-based APIs and return tagged data using XML or JSON, etc. Examples include Wikipedia, Twitter and Facebook.

• **Text cleaning and storage tools**—tools for cleaning and storing textual data. Google Refine and Data Wrangle rare examples for data cleaning.
• **Text analysis tools**—individual or libraries of tools for analyzing social media data once it has been scraped and cleaned. These are mainly natural language processing, analysis and classification tools, which are explained below.

- **Transformation tools**—simple tools that can transform textual input data into tables, maps, charts (line, pie, scatter, bar, etc.), timeline or even motion (animation over timeline), such as Google Fusion Tables, Zoho Reports, Tableau Public or IBM's Many Eyes.
- **Analysis tools**—more advanced analytics tools for analyzing social data, identifying connections and building networks, such as Gephi (open source) or the Excel plug-in NodeXL.

• **Social media platforms**—environments that provide comprehensive social media data and libraries of tools for analytics. Examples include: Thomson Reuters Machine Readable News, Radian 6 and Lexalytics.

- **Social network media platforms**—platforms that provide data mining and analytics on Twitter, Facebook and a wide range of other social network media sources.
- **News platforms**—platforms such as Thomson Reuters providing commercial news archives/feeds and associated analytics.

5. 1 Social media methodology and critique

The two major impediments to using social media for academic research are firstly access to comprehensive data sets and secondly tools that allow 'deep' data analysis without the need to be able to program in a language such as Java. The majority of social media resources are commercial and companies are naturally trying to monetize their data. As discussed, it is important that researchers have access to open-source 'big' (social media) data sets and facilities for experimentation. Otherwise, social media research could become the exclusive domain of major companies, government agencies and a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Recently, there has been a modest response, as Twitter and Gnip are piloting a new program for data access, starting with 5 all-access data grants to select applicants.

5.2. Methodology

Research requirements can be grouped into: data, analytics and facilities.

5.2.1 Data

Researchers need online access to historic and real-time social media data, especially the principal sources, to conduct world-leading research:

• **Social network media**—access to comprehensive historic data sets and also real-time access to sources, possibly with a (15 min) time delay, as with Thomson Reuters and Bloomberg financial data.

• **News data**—access to historic data and real-time news data sets, possibly through the concept of 'educational data licenses' (cf. software license).

• **Public data**—access to scraped and archived important public data; available through RSS feeds, blogs or open government databases.

• **Programmable interfaces**—researchers also need access to simple application programming interfaces (APIs) to scrape and store other available data sources that may not be automatically collected.

5.2.2 Analytics

Currently, social media data is typically either available via simple general routines or require the researcher to program their analytics in a language such as MATLAB, Java or Python. As discussed above, researchers require:

- **Analytics dashboards**—non-programming interfaces are required for giving what might be termed as 'deep' access to 'raw' data.
- **Holistic data analysis**—tools are required for combining (and conducting analytics across) multiple social media and other data sets.
- **Data visualization**—researchers also require visualization tools whereby information that has been abstracted can be visualized in some schematic form with the goal of communicating information clearly and effectively through graphical means.

5.2.3 Facilities

Lastly, the sheer volume of social media data being generated argues for national and international facilities to be established to support social media research (cf. Wharton Research Data Services https://wrds-web.wharton.upenn.edu):

• **Data storage**—the volume of social media data, current and projected, is beyond most individual universities and hence needs to be addressed at a national science foundation level. Storage is required both for principal data sources (e.g., Twitter), but also for sources collected by individual projects and archived for future use by other researchers.

• **Computational facility**—remotely accessible computational
facilities are also required for: a) protecting access to the stored data; b) hosting the analytics and visualization tools; and c) providing computational resources such as grids and GPUs required for processing the data at the facility rather than transmitting it across a network.

5.2.4 Critique

Much needs to be done to support social media research. As discussed, the majority of current social media resources is commercial, expensive and difficult for academics obtain full access.

5.2.5 Data

In general, access to important sources of social media data is frequently restricted and full commercial access is expensive.

• **Siloed data**—most data sources (e.g., Twitter) have inherently isolated information making it difficult to combine with other data sources.

• **Holistic data**—in contrast, researchers are increasingly interested in accessing, storing and combining novel data sources: social media data, real-time financial market & customer data and geospatial data for analysis. This is currently extremely difficult to do even for Computer Science departments.

5.2.6 Analytics

Analytical tools provided by vendors are often tied to a single data set, maybe limited in analytical capability, and data charges make them expensive to use.

5.2.6 Facilities

There are an increasing number of powerful commercial platforms, such as the ones supplied by SAS and Thomson Reuters, but the charges are largely prohibitive for academic research. Either comparable facilities need to be provided by national science foundations or vendors need to be persuaded to introduce the concept of an 'educational license.'

5.3 Social media data

Clearly, there is a large and increasing number of commercial) services providing access to social networking media (e.g., Twitter, Facebook and Wikipedia) and news services (e.g., Thomson Reuters Machine Readable News). Equivalent major academic services are scarce. We start by discussing types of data and formats produced by these services.

5.3.1 Types of data

Although we focus on social media, as discussed, researchers are continually finding new and innovative sources of data to bring together and analyze. So when considering textual data analysis, we should consider multiple sources (e.g., Social networking media, RSS feeds, blogs and news) supplemented by numeric (financial) data, telecoms data, geospatial data and potentially speech and video data. Using multiple data sources is certainly the future of analytics. Broadly, data subdivides into:

• **Historic data sets**—previously accumulated and stored social/news, financial and economic data.

• **Real-time feeds**—live data feeds from streamed social media, news services, financial exchanges, telecoms services, GPS devices and speech.
And into:

• **Raw data**—unprocessed computer data straight from sources that may contain errors or may be unanalyzed.

• **Cleaned data**—correction or removal of erroneous (dirty) data caused by disparities, keying mistakes, missing bits, outliers, etc.

• **Value-added data**—data that has been cleaned, analyzed, tagged and augmented with knowledge.

5.3.2 Text data formats

The four most common formats used to markup text are: HTML, XML, JSON and CSV.

• HTML—Hyper Text Markup Language (HTML) as well-known is the markup language for web pages and other information that can be viewed in a web browser. HTML consists of HTML elements, which include tags enclosed in angle brackets (e.g., \div[), within the content of the web page.

• XML—Extensible Markup Language (XML)—the markup language for structuring textual data using \tag[…\\tag[to define elements.

• JSON—JavaScript Object Notation (JSON) is a text based open standard designed for human-readable data interchange and is derived from JavaScript.

• CSV—a comma-separated values (CSV) file contains the values in a table as a series of ASCII text lines organized such that each column value is separated by a comma from the next column's value and each row starts a new line.

For completeness, HTML and XML are so-called markup languages (markup and content) that define a set of simple syntactic rules for encoding documents in a format both human readable and machine readable. A markup comprises start-tags (e.g., \tag[), content text and endtags (e.g.,\/tag[).

Many feeds use JavaScript Object Notation (JSON), the light weight data-interchange format, based on a subset of the JavaScript Programming Language. JSON is a language-independent text format that uses conventions that are familiar to programmers of the C-family of languages, including C, C??, C#, Java, JavaScript, Perl, Python, and many others.

JSON's basic types are: Number, String, Boolean, Array (an ordered sequence of values, comma separated and enclosed in square brackets) and Object (an unordered collection of key: value pairs). The JSON format is illustrated in Fig. 1 for a query on the Twitter API on the string 'UCL,' which returns two 'text' results from the Twitter user 'uclnews.' Comma-separated values are not a single, well-defined format but rather refer to any text file that: (a) is plain text using a character set such as ASCII, Unicode or EBCDIC; (b) consists of text records (e.g., one record per line);(c) with records divided into fields separated by delimiters (e.g., comma, semicolon and tab); and (d) where every record has the same sequence of fields.

VI. SOCIAL MEDIA PROVIDERS

Social media data resources broadly subdivide into those providing:

- Freely available databases—repositories that can be freely downloaded, e.g., Wikipedia (http://dumps.wikimedia.org) and the Enron e-mail data set available via http://www.cs.cmu.edu/*enron/.

- Data access via tools—sources that provide controlled access to their social media data via dedicated tools, both to facilitate easy interrogation and also to stop users 'sucking' all the data from the repository. An example is Google's Trends. These further subdivided into:
- Free sources—repositories that are freely accessible, but the tools protect or may limit access to the 'raw' data in the repository, such as the range of tools provided by Google.

- Commercial sources—data resellers that charge for access to their social media data. Gnip and DataSift provide commercial access to Twitter data through a partnership, and Thomson Reuters to news data.

- Data access via APIs—social media data repositories providing programmable HTTP-based access to the data via APIs (e.g., Twitter, Facebook and Wikipedia).

VII. SOCIAL MEDIA ANALYTICS TECHNIQUES

As discussed, opinion mining (or sentiment analysis) is an attempt to take advantage of the vast amounts of user-generated text and news content online. One of the primary characteristics of such content is its textual disorder and high diversity. Here, natural language processing, computational linguistics and text analytics are deployed to identify and extract subjective information from source text. The general aim is to determine the attitude of a writer (or speaker) with respect to some topic or the overall contextual polarity of a document.

VIII. SENTIMENT ANALYSIS

Sentiment is about mining attitudes, emotions, feelings—it is subjective impressions rather than facts. Generally speaking, sentiment analysis aims to determine the attitude expressed by the text writer or speaker with respect to the topic or the overall contextual polarity of a document (Mejova 2009). Pang and Lee (2008) provide a thorough documentation on the fundamentals and approaches of sentiment classification and extraction, including sentiment polarity, degrees of positivity, subjectivity detection, opinion identification, non-factual information, term presence versus frequency, POS (parts of speech), syntax, negation, topic-oriented features and term-based features beyond term unigrams.

8.1.1 Sentiment classification

Sentiment analysis divides into specific subtasks:

• Sentiment context—to extract opinion, one needs to know the 'context' of the text, which can vary significantly from specialist review portals/feeds to general forums where opinions can cover a spectrum of topics (Westerski 2008).

• Sentiment level—text analytics can be conducted at the document, sentence or attribute level. Sentiment subjectivity—deciding whether a given text expresses an opinion or is factual (i.e., without expressing a positive/negative opinion).

• Sentiment orientation/polarity—deciding whether an opinion in a text is positive, neutral or negative.

• Sentiment strength—deciding the 'strength' of an opinion in a text: weak, mild or strong. Perhaps, the most difficult analysis is identifying sentiment orientation/polarity and strength—positive (wonderful, elegant, amazing, cool), neutral (fine, ok) and negative (horrible, disgusting, poor, flakey, sucks) due to slang.

A popular approach is to assign orientation/polarity scores (?1, 0, -1) to all words: positive opinion (?1), neutral opinion (0) and negative opinion (-1). The overall orientation/polarity score of the text is the sum of orientation scores of all 'opinion' words found. However, there are various potential problems in this simplistic approach, such as negation (e.g., there is nothing I hate about this product). One method of estimating sentiment orientation / polarity of the text is point wise mutual information (PMI) a measure of association used in information theory and statistics.

8.2.2 Social media analytics tools

Opinion mining tools are crowded with (commercial) providers, most of which are skewed toward sentiment analysis of customer feedback about products and services. Fortunately, there is a vast spectrum of tools for textual analysis ranging from simple open-source tools to libraries, multi-function commercial toolkits and platforms. This section focuses on individual tools and toolkits for scraping, cleaning and analytics, and the next chapter looks at what we call social media platforms that provide both archive data and real-time feeds, and as well as sophisticated analytics tools.

8.3 Social media monitoring tools

Social media monitoring tools are sentiment analysis tools for tracking and measuring what people are saying (typically) about a company or its products, or any topic across the web's social media landscape.

In the area of social media monitoring examples include: Social Mention, (http://socialmention.com/), which provides social media alerts similarly to Google Alerts; Amplified Analytics (http://www.amplifiedanalytics.com/), which focuses on product reviews and marketing information; Lithium Social Media Monitoring; and Trackur, which is an online reputation monitoring tool that tracks what is being said on the Internet. Google also provides a few useful free tools. Google Trends shows how often a particular search-term input
compares to the total search volume. Another tool built around Google Search is Google Alerts—a content change detection tool that provides notifications automatically. Google also acquired FeedBurner—an RSS feeds management— in 2007.

IX. CONCLUSION AND FUTURE WORK

Now, computer industry accepts Big Data as a new challenge for all types of machine automated systems. There are many issues in storage, management, and retrieval of data known as Big Data. The main problem is how we can use this data for increasing business and improvement in living standard of people. In this paper we are discussing the issues, challenges, application as well as proposing some actionable insight for Big Data. It will motivate researchers for finding knowledge from the big amount of data available in different forms in different areas. This paper presents the fundamental concepts of Big Data. This paper surveys some of the social media software tools, and for completeness introduced social media scraping, data cleaning and sentiment analysis. These concepts include the increase in data, the progressive demand for HDDs, and the role of Big Data in the current environment of enterprise and technology numbered with separate brackets. When citing a section in a

REFERENCES

[1]. Web content available on the link: —http://www.sas.com/en_us/insights/big-data/what-is-big-data.html‖ on the dated: 16-08-2015
[2]. Web content available on the link: —http://www-01.ibm.com/software/data/bigdata/what-is-big-data. html‖ on the dated: 16-08-2015
[3]. Web content available on the link: —http://www.dqindia.com/8-innovative-examples-of-big-data-usage-in-india/‖ on the dated: 16-08-2015
[4]. Web content available on the link: —http://searchbusinessanalytics.techtarget.com/definition/big-dataanalytics‖ on the dated: 16-08-2015
[5]. Web content available on the link: —https://www.statsoft.com/textbook/text-mining‖ on the dated: 16- 08-2015
[6]. Web content available on the link: —https://www.predictiveanalyticstoday.com/text-analytics‖ on the dated: 16-08-2015
[7]. Web content available on the link: —https://gigaom.com/2014/01/24/why-video-is-the-next-big-thingin- big-data/‖ on the dated: 16-08-2015
[8]. Web content available on the link: —http://searchbusinessanalytics.techtarget.com/definition/socialmedia - analytics‖ on the dated: 16-08-2015
[9]. Web content available on the link: ‖http://en.wikipedia.org/wiki/E-commerce_in_India#cite_note-
[10]. Online_shopping_touched_new_heights_in_India_in_2012-1‖ on the dated: 16-08-2015.
[11]. Amir Gandomi , Murtaza Haider, —Beyond the hype: Big data concepts, methods, and analytics‖, International Journal of Information Management 35 (2015) 137–144, journal homepage: www.elsevier.com/locate/ijinfomgt
[12]. Radhakrishnan, Ajay Divakaran and Paris Smaragdis, —AUDIO ANALYSIS FOR SURVEILLANCE APPLICATIONS‖, 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 16-19, 2005, New Paltz, NY.
[13]. C.L. Philip Chen, Chun-Yang Zhang, 2014, —Data-intensive applications, challenges, techniques and

technologies: A survey on Big Data‖, Contents lists available at Science Direct Information Sciences, 275 (2014) 314–347
[14]. Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, 2013, ―Big Data: Issues and Challenges Moving Forward‖, 2013 46th Hawaii International Conference on System Scien 1530-1605/12, 2012 IEEE,
[15]. Web content available on the link: ―http://analyticstraining.com/2014/the-power-of-social-mediaanalytics/‖on the dated: 17-10-2015
[16]. Edmon Begoli, James Horey, 2012, ―Design Principles for Effective Knowledge Discovery from Big Data‖, 2012 Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture, 978-0-7695-4827-2/12, IEEE
[17]. Yang Song, Gabriel Alatorre, Nagapramod Mandagere, and Aameek Singh, 2013, ―IEEE International Congress on Big Data‖, 978-0-7695-5006-0/13, IEEE
[18]. "Social Media and Big Data" POSTNOTE house of parliament No. 460 March 2014
[19]. Botan I et al. (2010) SECRET: a model for analysis of the execution semantics of stream processing systems. Proc VLDB Endow 3(1–2):232–243
[20]. Salathe´ M et al. (2012) Digital epidemiology. PLoS Comput Biol 8(7):1–5
[21]. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(3):1–8
[22]. Chandramouli B et al (2010) Data stream management systems for computational finance. IEEE Comput 43(12):45–52
[23]. Chandrasekar C, Kowsalya N (2011) Implementation of MapReduce Algorithm and Nutch Distributed File System in Nutch. Int J Comput Appl 1:6–11
[24]. Cioffi-Revilla C (2010) Computational social science. Wiley Interdiscip Rev Comput Statistics 2(3):259–271
[25]. Galas M, Brown D, Treleaven P (2012) A computational social science environment for financial/economic experiments. In: Proceedings of the Computational Social Science Society of the Americas, vol 1, pp 1–13
[26]. Hebrail G (2008) Data stream management and mining. In: Fogelman- Soulie´ F, Perrotta D, Piskorski J, Steinberger R (eds) Mining Massive Data Sets for Security. IOS Press, pp 89–102
[27]. Hirudkar AM, Sherekar SS (2013) Comparative analysis of data mining tools and techniques for evaluating performance of database system. Int J Comput Sci Appl 6(2):232–237
[28]. Kaplan AM (2012) If you love something, let it go mobile: mobile marketing and mobile social media 4x4. Bus Horiz 55(2):129–139
[29]. Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 53(1):59–68
[30]. Karabulut Y (2013) Can Facebook predict stock market activity? SSRN eLibrary, pp 1–58. http://ssrn.com/abstract=2017099 or http://dx.doi.org/10.2139/ssrn.2017099. Accessed 2 Feb 2014
[31]. Khan A, Baharudin B, Lee LH, Khan K (2010) A review of machine learning algorithms for text-documents classification. J Adv Inf Technol 1(1):4–20
[32]. Kobayashi M, Takeda K (2000) Information retrieval on the web. ACM Comput Surv CSUR 32(2):144–173
[33]. Lazer D et al (2009) Computational social science. Science 323:721–723
[34]. Lerman K, Gilder A, Dredze M, Pereira F (2008) Reading the markets: forecasting public opinion of political candidates by news analysis. In: Proceedings of the 22nd international conference on computational linguistics 1:473–480
[35]. MapReduce (2011) What is MapReduce?. http://www.mapreduce. org/what-is-mapreduce.php. Accessed 31 Jan 2014
[36]. Mejova Y (2009) Sentiment analysis: an overview, pp 1–34. http://
[37]. www.academia.edu/291678/Sentiment_Analysis_An_Overview. Accessed 4 Nov 2013
[38]. Murphy KP (2006) Naive Bayes classifiers. University of British Columbia, pp 1–8. http://www.ic.unicamp.br/*rocha/teaching/2011s1/mc906/aulas/naivebayes.pdf
[39]. Murphy KP (2012) Machine learning: a probabilistic perspective. In: Chapter 1: Introduction. MIT Press, pp 1–26
[40]. Narang RK (2009) Inside the black box. Hoboken, New Jersey Nuti G, Mirghaemi M, Treleaven P, Yingsaeree C (2011) Algorithmic trading. IEEE Comput 44(11):61–69
[41]. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135
[42]. SAS Institute Inc (2013) SAS sentiment analysis factsheet. http://www. sas.com/resources/factsheet/sas-sentiment-analysis-factsheet.pdf.
[43]. Accessed 6 Dec 2013 Teufl P, Payer U, Lackner G (2010) From NLP (natural language processing) to MLP (machine language processing). In: KotenkoI, Skormin V (eds) Computer network security, Springer, Berlin Heidelberg, pp 256–269
[44]. Thomson Reuters (2010). Thomson Reuters news analytics. http:// thomsonreuters.com/products/financial-risk/01_255/News_Analy tics_-_Product_Brochure-_Oct_2010_1_.pdf. Accessed 1 Oct 2013
[45]. Thomson Reuters (2012) Thomson Reuters machine readable news. http://thomsonreuters.com/products/financial-risk/01_255/TR_ MRN_Overview_10Jan2012.pdf. Accessed 5 Dec 2013
[46]. Thomson Reuters (2012) Thomson Reuters MarketPsych Indices. http://thomsonreuters.com/products/financial-risk/01_255/TRMI_flyer_2012.pdf. Accessed 7 Dec 2013
[47]. Thomson Reuters (2012) Thomson Reuters news analytics for internet news and social media. http://thomsonreuters.com/business-unit/financial/eurozone/112408/news_analytics_and_social_media. Accessed 7 Dec 2013
[48]. Thomson Reuters (2013) Machine readable news. http://thomsonreuters.com/machine-readable-news/?subsector=thomson-reuters-elektron. Accessed 18 Dec 2013

[49]. Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics pp. 417–424
[50]. Vaswani V (2011) Hook into Wikipedia information using PHP and the MediaWiki API. http://www.ibm.com/developerworks/web/library/x-phpwikipedia/index.html. Accessed 21 Dec 2012
[51]. Westerski A (2008) Sentiment analysis: introduction and the state of the art overview. Universidad Politecnica de Madrid, Spain, pp 1–9. http://www.adamwesterski.com/wpcontent/files/docsCursos/sentimentA_doc_TLAW.pdf. Accessed 14 Aug 2013
[52]. Wikimedia Foundation (2014) Wikipedia:Database download. http://en.wikipedia.org/wiki/Wikipedia:Database_download. Accessed 18 Apr 2014
[53]. Wolfram SMA (2010) Modelling the stock market using Twitter. Dissertation Master of Science thesis, School of Informatics