Multimedia Social Big Data: Mining

Akshi Kumar, Saurabh Raj Sangwan and Anand Nayyar

Abstract The rapid evolution and adoption of the SMAC (Social media, Mobile, Analytics and Cloud) technology paradigm, has generated massive volumes of human-centric, real-time, multimodal, heterogeneous data. Human-sourced information from social networks, process-mediated data from business systems and machine-generated data from Internet-of-Things are the three primary sources of big data which define the richness and scale of multimedia content available. With the proliferation of social networks (Twitter, Tumblr, Google+, Facebook, Instagram, Snapchat, YouTube, etc.), the user can post and share all kinds of multimedia content (text, image, audio, video) in the social setting using the Internet without much knowledge about the Web's client-server architecture and network topology. This proffer novel opportunities and challenges to leverage high-diversity multimedia data in concurrence to the huge amount of social data. In recent years, multimedia analytics as a technology-based solution has attracted a lot of attention by both researchers and practitioners. The mining opportunities to analyze, model and discover knowledge from the social web applications/services are not restricted to the text-based big data, but extend to the partially unknown complex structures of image, audio and video. Interestingly, the big data is estimated to be 90% unstructured further, making it crucial to tap and analyze information using contemporary tools. The work presented is an extensive and organized overview of the multimedia social big data mining and applications. A comprehensive coverage of the taxonomy, types and techniques of Multimedia Social Big Data mining is put forward. A SWOT Analysis is done to understand the feasibility and scope of social multimedia content and big data analytics is also illustrated. Recent applications and suitable directions for

A. Kumar · S. R. Sangwan
Department of Computer Science & Engineering, Delhi Technological University,
New Delhi, India
e-mail: akshikumar@dce.ac.in

S. R. Sangwan
e-mail: saurabhsangwan2610@gmail.com

A. Nayyar (✉)
Graduate School, Duy Tan University, Da Nang, Vietnam
e-mail: anandnayyar@duytan.edu.vn

© Springer Nature Singapore Pte Ltd. 2020

S. Tanwar et al. (eds.), *Multimedia Big Data Computing for IoT Applications*, Intelligent Systems Reference Library 163,
https://doi.org/10.1007/978-981-13-8759-3_11

future research have been identified which validate and endorse this correlation of multimedia to big data for mining social data.

Keywords Big data · Social data · Web mining

1 Introduction to Multimedia Social Big Data Mining

The unnceasing surge of the World Wide Web and the acquaintance of the people around the world with the Internet create all the required prerequisites for a wide-ranging adaptation of the elementary Internet as a general medium for exchange of information, where any user can become a contributor. This new collaborative Web (called Web 2.0) resiliently defines the techno-social system which augments human cognition, communication, and co-operation; where cognition is the necessary prerequisite to communicate and the precondition to co-operate. In other words, cooperation needs communication and communication needs cognition. Such complex social big data (human-sourced, process-mediated and device generated) calls for cross disciplinary research from data mining, machine learning, pervasive and ubiquitous computing, networking, and computational social science.

The term "Big Data" is of huge importance today for researchers and practitioners alike. It is defined as '*high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making'*. Big Data is used to denote a collection of data sets which are too large and complex to handle and process using conventional data processing tools and applications. The size of data sets cannot be handled by common software tools used to for capturing, curating, managing and processing data within an acceptable time limit. The characteristics of Big Data are described by 7 V's given in the following Table 1.

Big Data is a trending set of techniques that demand new ways of consolidation of various methods to uncover hidden information from the massive and complex raw supply of data. User-generated content on the Web has been established as a type of big data and thus the discussion on Big data is inevitable in any description of the evolution of Web and its growth. Following are the types of Big Data that have been identified across the literature:

- **Social Networks (human-sourced information)**: Almost all of the data and information generated by people today is digitized and can be stored on any device from PCs to social websites. Such data is usually not structured and not governed by anyone. Data generated from social media platforms like Facebook, Blog posts and comments, Personal documents, Pictures from Instagram, Flickr, etc., and Videos from YouTube, E-mail, Web search results and content from mobile phones: text messages amongst others.
- **Traditional Business systems (process-mediated data)**: Business events like customer registration, product-manufacturing, placing and taking orders, etc. are

Table 1 Big-data Characteristics

Volume Scale of data	It determines the amount of data that flows in, which can be stored and originated further. Depending on the amount of data that is stored, it is decided whether it can fall in the category of "big data" or not
Variety Different forms of data	Different type and various sources of data are specified including both structured and unstructured data. For e.g. documents, emails, images, videos, audio etc.
Velocity Analysis of Streaming data	It is the aspect which deals with the pace of the data in motion and its analysis where the content flow is assumed to be continuous and immense. Apart from the consideration of the speed of the input, it also contemplates the celerity of the generation of useful information from it
Veracity Uncertainty of data	This characteristic of big data deals with the primary issue of reliability and whether the analysis of data is being accurately done, so that eventually there is a production of credible and quality solutions
Variability Contextual meaning of data	It refers to the inconsistency of the information that is stored. In simpler words, it deals with rapidly changing and alternative meanings that are associated with the data
Visualization Way to represent information/data	Visualization happens to make one of the crucial characteristics of big data because all the data that is being stored used as an input and generated as a result, needs to be sorted and viewed in a manner that is easy to read and comprehend
Value Cost of using data	It deals with the practice of retrieving the usefulness of the data. It is perceived that data in its original self won't be valuable at all. Under analysis, how data is turned into knowledge and information is what the "value" characteristic deals with

recorded and monitored by these. Such data include data from transactions, tables of references, relations, etc. and is quite structured. Data generated by Public organizations like Medical records and Data from Commercial transactions, Bank and stock records, E-commerce and credit/debit cards etc. are some examples.

- **Internet of Things (machine-generated data)**: This includes data generated by the huge number of sensors and machines that are used for measuring and recording the events in the real world. The data from sensors is well structured as it is machine-generated, and range from simple sensor records to complex computer logs. Data from fixed sensors like home automation, weather/pollution sensors, traffic sensors/webcam and security/surveillance videos/images, data from mobile sensors (tracking) such as mobile phone location, cars and satellite images and data from computer systems such as logs and web logs are some examples.

Business innovation and social intelligence pave the way to a future in which smart factories, intelligent machines, networked processes and big data are brought together to foster industrial growth and shift the economics. Social media has emerged as a key player which provides a platform for expression and distribution of content in today's world. The primary intent of social networking sites is to create, professional, interest, relationship-based virtual communities, enabling stronger connections with everyone around the world. Social media content is one of the major data sections which have attracted people and organizations across the globe. This is primarily due to the omnipresence and coverage of social applications world-wide. Social Media applications allow users to share comments, opinions, ideas, and media with friends, family, businesses, and organizations [1]. The data contained in these comments, ideas, and media are valuable to many types of organizations. Moreover, social media is inherently an informal way of communication with all kinds of multimedia content.

Social media dynamics keep changing with respect to the increasing user base and user-activity which makes it a high dimensional, complex and fuzzy data space for analytical processing. Thus, social big data mining contributes significantly to the field of data analytics as a new technology-based solution to comprehend wider and deeper applications of social media data. The maturity and growth of data science and machine learning for real-time decision making are the key technology drivers supporting viable, competent, and an actionable business model.

Social Media is one of the largest contributors to big data. With the increasing number of social media platforms and the rapid increase in the number of users of these platforms, huge amounts of data are generated around the world by the minute. Facebook, Twitter, Instagram and YouTube are some of the prominent contributors to social media. Analyzing the data from these websites would not only improve decisions, but would also lead to cost reduction and new products and services better suited to the needs of the users. The following Fig. 1 depicts the role of social big-data analytics in a typical business setting.

The following Fig. 2 depicts the existence of multimedia social big data in a typical business setting.

Variety and velocity are two key terms associated with big data. There are various sources of big data. Clickstream data, data from various electronic devices like sensors, social media platforms like Facebook, Twitter, LinkedIn, Instagram, Google+ tremendously contribute a variety of data apart from text like images, videos and audios. Geo-spatial data sharing people's locations feeds from business companies, markets and even those related to government are increasing day by day. Web 2.0 including data from mobiles, e-commerce websites like Amazon, Flipkart, eBay, recommendations as well business and marketing data and data from Enterprise Resource Planning and Customer Relationship Management like bills and data about stocks and inventory are subsets that contribute towards big data.

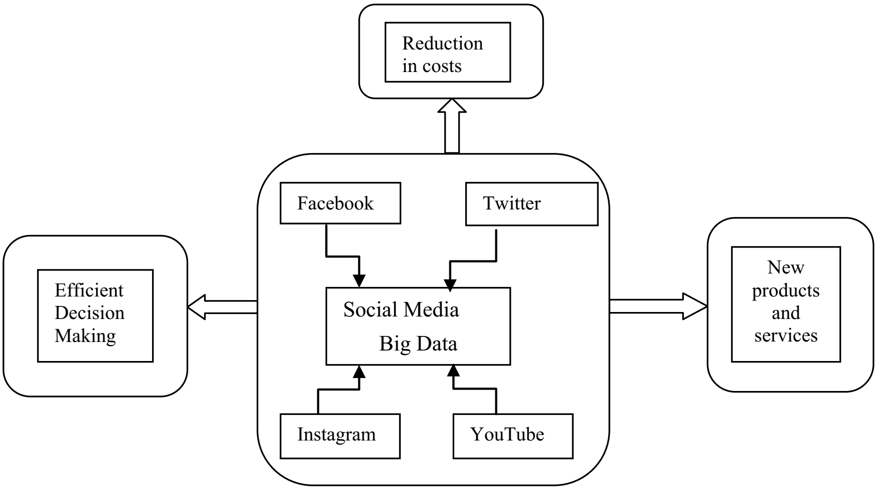

Fig. 1 Social big-data analytics in business

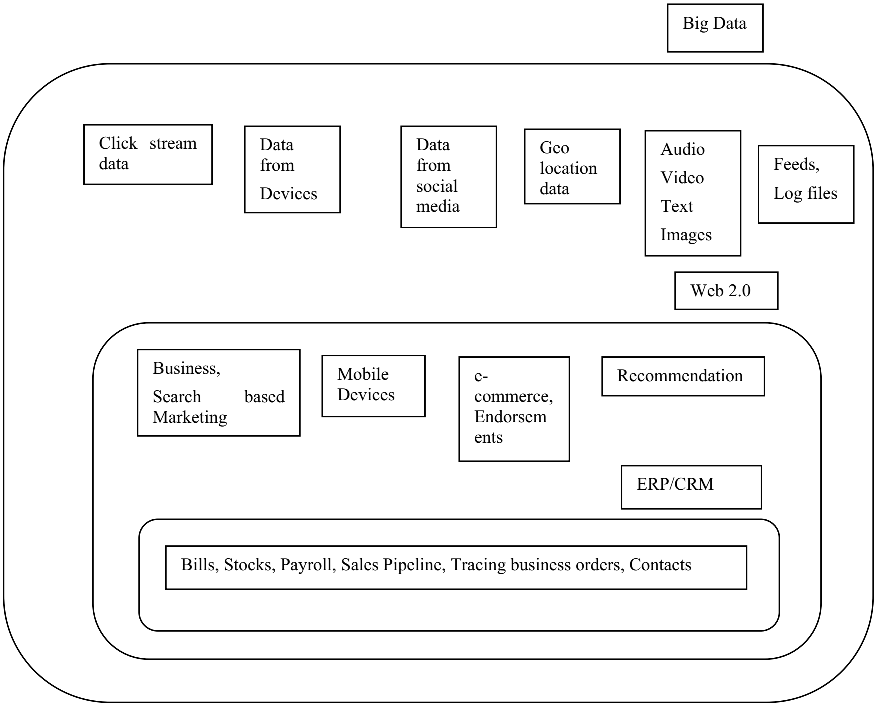

Fig. 2 Multimedia social big data in business

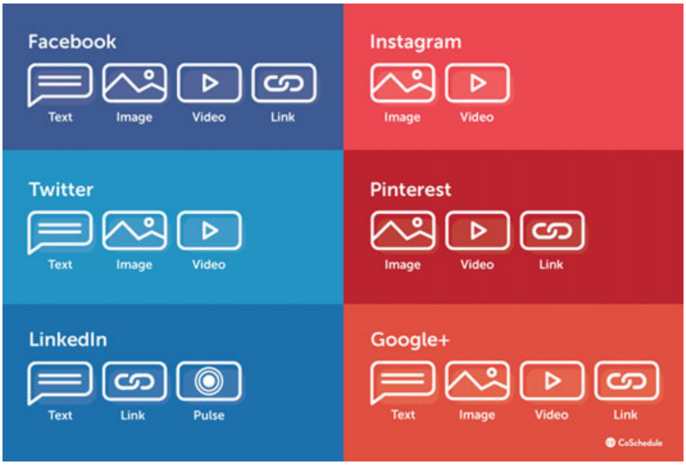

Fig. 3 Multimedia support by popular social networking sites

Recently, visual communication using images to express views, opinions, feelings, emotions and sentiments has increased tremendously on social platforms like Flickr, Instagram, Twitter, Tumblr, etc. [2–5]. Images are particularly powerful as they have cognition associated and visual experiences convey sentiments and emotions better. Consequently, the visual sentiment analysis has been of interest to researchers and it has been observed that deep learning techniques have outperformed the conventional machine learning techniques in analyzing the visual sentiment. Multimodal capabilities offered by popular social networking websites such as Facebook, Twitter, and Tumblr have further enabled mix of text and images in a variety of ways for better social engagement. The ascendant use of infographics, typographic-images, memes and GIFs in social feeds is a testimony to this. Text-driven analytics has been widely studied [6–9] and few pertinent studies which report visual analytics of images are available in literature [10–14]. Moreover, much of the reported work has analyzed a single modality data, whereas multiple modalities of text and image remain unexplored. The following Fig. 3 depicts the multimedia types supported by popular social networking sites.

The following Table 2 lists out the media types supported by popular social media platforms.

2 Process Model for Multimedia Social Big Data Mining

Big Data is a term used to describe data that is huge in amount and which keeps growing with time. Big Data consists of structured, unstructured and semi-structured data. This data can be used to track and mine information for analysis or research purpose.

Table 2 Popular social media characteristics

Social media platform	Multimedia supported	Characteristics
Facebook	Text, images, videos, links	Timeline, wall, events, status, embed-in posts, social plug-ins
Twitter	Text, images, videos	Pinning tweets, advanced search, Twitter moments, customized tweet alerts
LinkedIn	Text, links, pulse	Keep in touch, Get help, search for jobs, hire new employees
Instagram	Images, videos	Live Videos, Stories, Push notifications, Filters
Pinterest	Images, videos, links	Article Pins, Ad groups, Lens, Shop the Look, Pinterest e-mail, Native Video, Cinematic Pins
Google+	Text, images, videos, links	Google+ circles, authorship, hangouts, communities, events, Insights, My Business

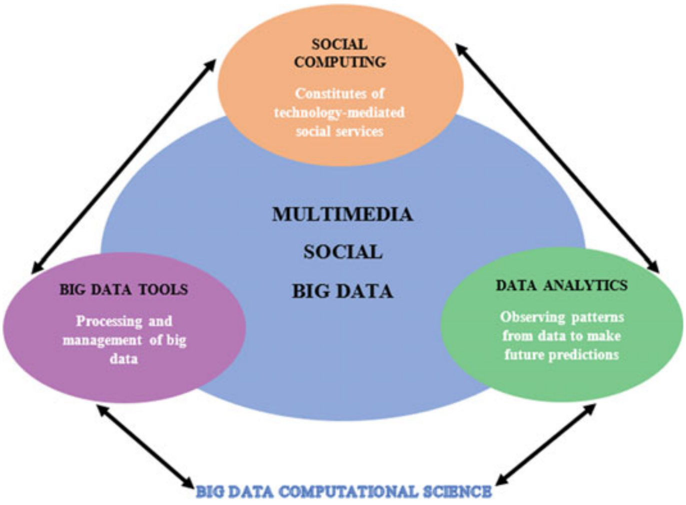
SOCIAL COMPUTING
Constitutes of technology-mediated social services
MULTIMEDIA SOCIAL BIG DATA
BIG DATA TOOLS
Processing and management of big data
DATA ANALYTICS
Observing patterns from data to make future predictions
BIG DATA COMPUTATIONAL SCIENCE

Fig. 4 Big data aspects

Big data and analytics together can not only be helpful in determining primary reasons for loss in businesses, but can also be used for analyzing trends in sales on the basis of customer buying history. It can also be helpful in determining fraudulent behavior and thus helps in reducing potential risks for organizations. Figure 4 represents Big data and its various aspects.

Big data contain complex and huge datasets thus making it difficult to process it using conventional tools and software. Multimedia social big data encompasses a variety of data and it is challenging to process and manage such data especially related to its volume and complexity. A few important challenges for multimedia big data are:

- Structuring the unstructured data.
- Understand and capture most important data eradicating irrelevant and redundant data.
- Storing the huge amount of data that is growing day by day.

Some of the tools used for storage and analysis of big data are Apache Hadoop; Microsoft HDInsight; Hive; Sqoop; PolyBase; Presto.

Another important aspect of social big data is Social Computing. Social Computing is a research domain that helps us in understanding social behaviors. It is concerned with the intersection of social behavior and computational systems and deals with the mechanisms through which people interact with computational systems [15]. Some examples where social computing has helped in research in recent times is given as follows:

- Researchers have looked at how people behave in various online communities (such as Subreddits, Yikyak, etc.).
- What kind of users visit a website with what goals and the persona a person holds in their online presence.
- Examining questions such as how and why people contribute user-generated content and how to design systems that better enable them to do so.

Some examples include collective intelligence, prediction markets, crowdsourcing markets etc. The multimedia big-data from the popular social websites in the form of videos, images, text, emails, are captured and store into various data repositories from where it can be shared among users. Various steps to be followed in the analysis of big data are represented in Fig. 5. and are described as follows:

- **Data Pre-processing**

Data pre-processing is done for cleaning and transforming the data collected in order to remove noise.

Data cleaning includes

– Noise reduction
– Missing value imputation
– Inconsistencies elimination

Examples of data cleaning include removal of URLs, removal of punctuations including hash-tag ‘#’, removal of repetitive characters, and extra spaces. Further, tokenization is done to obtain a sequence of characters followed by stopping and stemming. Stemming is done to reduce the derived words to their word stem, base or root form. Stop words like the, a, is, she, him, on, etc. are removed as they are used for structuring of sentence and are less influential in identifying sentiment of a sentence. Followed by data cleaning the data is transformed into a structured data format. Data transformation includes

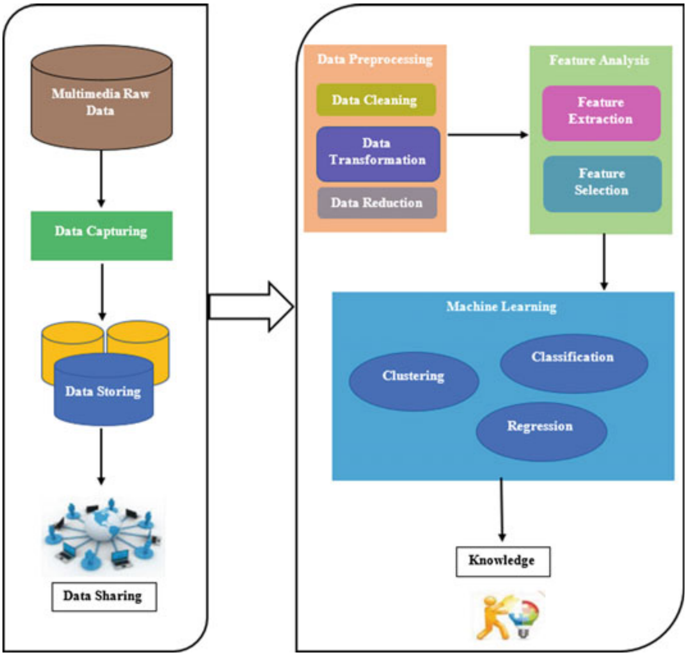

Fig. 5 Steps for analysis of big data

- Data normalization
- Data formatting
- Data aggregation

- **Feature Extraction**

This step identifies the characteristics of the datasets that are specifically useful in order to achieve our aim for example in detecting sentiments. There are different categories of features to be extracted namely, lexical, syntactic and semantic features which come under the category of morphological features. Other than these there are also frequent features and implicit features in the data.

- **Feature Selection**

Feature selection is the process of selecting a subset of features from the original set of features forming patterns in a given dataset [16]. Feature selection is done to reduce the size of problem for learning algorithms which may improve classification accuracy due to reduction in computation requirement as well as increase the speed of the classification task as the size of data to train the classifier is reduced. Feature selection techniques can be classified into two categories: filter approaches and wrapper approaches. The key difference between the two algorithms is that filter algorithms select the feature subset prior to the application of any classification algorithm, i.e. they are classification independent. The filter approach eliminates the less important features by using statistical properties of features. Wrapper algorithms select the features according to the accuracy of the training data and afterwards learn and test the classification model using the test data. Generally, for the implementation of wrapper methods learning algorithm and the performance criteria are defined.

- **Machine Learning**

Machine learning focuses on the development of computer programs that can access data and use for learning. Machine learning algorithms are broadly classified into Supervised, Unsupervised, and Semi-supervised [17]. Regression and Classification come under Supervised learning (answer for all the feature points are mapped) and Clustering comes under unsupervised learning (answer will not be given for the points).

- Regression—If the prediction value tends to be a continuous value, then it falls under Regression type problem in machine learning.
- Clustering includes grouping a set of points to the given number of clusters.
- Classification—If the prediction value tends to be categorized like yes/no, positive/negative, etc., then it falls under classification type problem in machine learning.

3 SWOT Analysis of Adopting Multimedia Social Big Data Mining in Real-Time Applications

The following Fig. 6 illustrates a typical lifecycle of data.

SWOT analysis can be defined as a model or technique which is used to figure out the strengths, weaknesses, opportunities and threats to a particular person, project or product. It lays the foundation stone for a project as it shows all the positives and negatives in advance and therefore, plays a crucial role in decision making. As the name implies, SWOT analysis for an entity deals with four factors:

- Strengths: The traits of the entity which provide it an edge over others.
- Weaknesses: The traits of the entity which are most likely to prove as a dis-benefit to itself.
- Opportunities: The external elements that the entity can use in order to get maximum benefits.
- Threats: The external elements that can prove to be hazardous for the entity in the future.

Considering the importance of SWOT analysis, a SWOT Analysis is carried out to get to know the feasibility and scope of social multimedia content and big data mining [18]. The following sub-sections discuss the details:

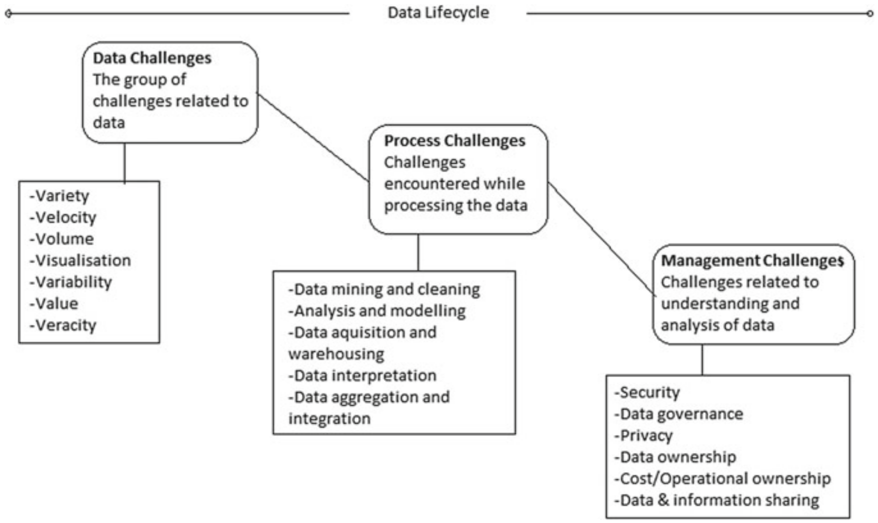
Data Lifecycle
Data Challenges
The group of challenges related to data
-Variety
-Velocity
-Volume
-Visualisation
-Variability
-Value
-Veracity
Process Challenges
Challenges encountered while processing the data
-Data mining and cleaning
-Analysis and modelling
-Data aquisition and warehousing
-Data interpretation
-Data aggregation and integration
Management Challenges
Challenges related to understanding and analysis of data
-Security
-Data governance
-Privacy
-Data ownership
-Cost/Operational ownership
-Data & information sharing

Fig. 6 Data lifecycle

3.1 Strengths

- **Enhance the performance of the business**

The major strength of multimedia social big data mining is that it upgrades the performance of businesses. This is due to the intelligent business systems that it provides. Every business can be analyzed with the use of big data to get to know what is working in its favor and what is working otherwise. The results can be used for modifying existing strategies for maximizing the profit.

- **Targeted users/customers**

Social multimedia content and big data mining provide the targeted customers to a business which will lead to more profit. This is done on the basis of past interests shown by a customer and other features of an individual's profile which may imply her/his liking for a particular product. This in turn benefits both, the business and the customer.

Also, in social networks like Twitter, Facebook, Quora, it is used to recommend news, users or stories to people based on their past choices and other factors derived from big data mining.

- **Speedy and precise analysis**

By using various machine learning techniques, the analysis of the Social multimedia content and big data has now become very fast. Moreover, if the most suitable technique is used for analyzing a particular problem, then the accuracy of the analysis is also laudable.

- **Improved Strategic decisions**

It is a well-known fact that one takes better decision when she/he is well aware of all the aspects of a situation. In this case also the social multimedia content and big data mining provides a deep insight in the situation which leads to outstanding decision making.

- **Customer/User satisfaction**

Due to targeted campaigning and better understanding of customer preferences provided by the big data, customer gets the required services easily without much effort. Also, in social networks, users get a better experience due to personalized content.

3.2 Weaknesses

- **Risk of wrong conclusions**

While analyzing social multimedia content and big data there are many factors that have to be kept in mind for an efficient and precise analysis, such as the expertise of team in the field they are dealing with. For example, if a team is analyzing a trend in the symptoms of a disease, then all members should have a basic understanding of medical science otherwise it may end up in very wrong predictions. Wrong conclusions are also derived when the data set is not of good quality.

- **Lack of professionals**

There is a need of small enterprises or people who are well efficient in carrying out the analysis of big data. The expertise in this field is a must. Therefore, the focus should be on training more and more individuals in this field so that we achieve the desired level of expertise.

- **Customer privacy at stake**

All the data we are talking about is customer specific. For analysis, one needs a well-defined and detailed data set which comes from the details of customers. This is one of the major weaknesses of social multimedia content and big data analysis.

- **Reduced data set may lead to incorrect analysis**

For making machine learning techniques more effective, we stress on reducing the size of data set, but sometimes this reduction might lead to an incorrect analysis as we may lose some important and deciding features.

3.3 Opportunities

- **Efficient tool for detecting rumor, bully, fake news etc.**

This is one of the most important opportunities in the field of analyzing social multimedia content and big data as it helps in making our environment safer. By using the machine learning techniques efficiently, we can come up with a solution to all these problems.

- **Education and health sector**

This is also a very important opportunity in this field as it touches millions of lives. By analyzing the data related to education sector, we can come to know about the performance of teachers. Also, whether the government education scheme is doing well or not. This will not only lead to better performance of students but also will increase the literacy rate. Similarly, in health sector social multimedia content and big data analysis can prove to be a game changer. Better, low cost and efficient treatment can be provided to everyone.

- **Fraud detection**

Fraud is a major problem every country is facing today. This is present in every sphere of life. Social multimedia content and big data analysis, if done properly, can lead to an efficient fraud detection system.

3.4 Threats

- **Privacy**

With this amount of access of an individual's data social multimedia content and big data analysis has to be very careful. As it may be opposed by people in the future.

- **Identity theft**

There is a very high chance of misuse of social multimedia content and big data mining. One of the main threats is about identity theft as with so much data available about an individual, one can easily access and misuse it for fraudulent practices.

- **Costs**

The costs involved in all these practices may rise up to a level where the cost-benefit ratio goes very low.

- **Social approval**

Social multimedia content and big data analysis may face a challenge in terms of social acceptance. This is due to the fact that it works in the public domain, uses people's information and recommends the results to them, i.e. totally public dependent. Therefore, social acceptance is a must which can be very difficult to achieve when people's privacy is at stake. The following Table 3 summarizes the SWOT Matrix.

Table 3 SWOT matrix

Strengths • Enhance the performance of the business • Targeted users/customers • Speedy and precise analysis • Speedy and precise analysis • Improved strategic decisions • Customer/User Satisfaction	**Weaknesses** • Risk of wrong conclusions • Customer privacy at stake • Lack of professionals • Reduced data set may lead to incorrect analysis
Opportunities • Efficient tool for detecting rumour, bully, fake news etc. • Education and health sector • Fraud detection	**Threats** • Privacy • Identity theft • Costs • Social approval

4 Techniques for Social Big Data Analytics

In this section various techniques which deal with social big data analytics are discussed. Data can be in any form such that text, audio, image or video. Different techniques are used for analyzing different types of data. Generally same methods are translated for different forms of data. The target of social big data analysis is to get to know people's behavior so that personalized content can be created for every individual. Therefore, the techniques are developed to get a deeper insight into how data is flowing or how is this data related to user's preference or what are the trending issues or whether a story is fake, rumour, bully, sarcasm, irony or genuine. The data under observation can be of any kind. Therefore, we need to have different kinds of techniques for different kinds of data. Broadly, this heterogeneity of data can be classified in four forms: Textual data, Audio data, Image data, Video data. Sometimes, when the data has more than one type in it, then we merge the techniques mentioned in this section for better analysis. Various techniques for different kinds of data are discussed in this section.

4.1 Text Analytics

In most of the cases we deal with textual data. This is because of the reason that mostly humans express their views either by speaking or writing. So, this is the basic type of data. Techniques used here are then translated into different forms so as to support other kinds of data. Text analysis mainly deals with the extraction of meaningful and structured data out of unstructured data. The main focus of text analytics is to extract the required information from textual data such as comments, stories, news, tweets, blogs, survey results, logs etc. The accurate analysis of textual data leads to a perfect decision making. Text analytics techniques are discussed below:

Information Extraction
This technique deals with the structuring of a huge unstructured data [19]. For example, one gets to understand the relationships between various entities are using this method. There are two methods which tell the correctness of an information extraction algorithm:

- Precision
- Recall

One of the major tasks of information extraction is to classify the data into entities, i.e. recognizing the class/entity to which a particular entry belongs. The techniques which are used for this classification of data entries into entities and then establishing relationships between these entities based on data are:

- Supervised machine learning techniques
- Semi-supervised machine learning techniques
- Unsupervised machine learning techniques

Supervised machine learning techniques includes Support Vector Machine (SVM), Hidden Markov models, decision tree, naive Bayes, k-nearest neighbor algorithm. Examples of semi-supervised machine learning techniques are bootstrapping, snowball, etc. Clustering algorithms are the examples of unsupervised machine learning techniques.

4.2 Audio Analytics

As the name suggests, audio analytics means analyzing the data which is in audio form to derive the required information from these audio signals. There are many applications of audio analytics. For example, in health care: Proper analysis of an infant's voice can tell his/her health status. Another example is in call centers, where audio/speech analysis is done to improve the quality of their service. The main techniques used for audio analysis are listed below [20].

- Approach based on phonetics

In this technique phonetics is used to classify the input audio. First, the system converts the input audio into a sequence of phonemes. Then the system searches for the output based on the labeling of phonemes.

- LVCSR

LVCSR stands for Large Vocabulary Continuous Speech Recognition. In this approach first step is to match sound to words using various algorithms like ASR (Automatic Speech recognition). The system matches the closest option if it fails to find a perfect match. On the basis of the output an indexed file is maintained, which provides the sequence in which words were spoken in the audio input. In the second step the required information is extracted from the indexed file obtained in step 1.

4.3 Image Analytics

Image analytics deals with the extraction of meaningful information from a data set composed of images. This is done using various image processing techniques. The major applications of image analytics are facial recognition, movement analysis. Nowadays, most of the unstructured data available is in image or video form. So, it becomes very important to come up with effective techniques for image analysis. For image analytics the methods or techniques used are basically the translated versions of basic machine learning techniques such as recurrent neural networks, convolution neural networks, deep learning algorithms etc. There is still a huge scope of research in this field. With increase in internet services and devices, more and more people and applications are contributing to data nowadays. To cope up with this speed, we need efficient techniques and infrastructure for better and accurate analysis of data in the form of an image.

4.4 Video Analytics

Video analytics is the most challenging one among all. It is mainly due to its size. We can get an idea of this problem with the fact that two second of an HD video is equivalent to 4000 pages of text, in terms of size. So much of data which is very large in size, due to its nature, is being generated in video form [21]. For example, CCTV footages, a camera in everyone's hand, videos uploaded to YouTube and other sites, etc. For analysis the basic machine learning techniques are modified for video analysis. These techniques also deal with the reduction in frame rate, resolution of images for easier analysis, but this reduction may lead to a less accurate prediction. Here also, there is a huge scope of improvement in techniques. A better database management system is also the need of the hour for analyzing data in the form of video. This will also reduce the training time.

5 Applications of Multimedia Social Big Data Mining

5.1 Trust and Information Veracity Analysis of Social Media Data

The web and social media have become an integral part of the daily lives of the general mass of people for the past decade. Social network platforms like Facebook and microblogging sites like Twitter have become ubiquitous. Given how readily accessible these platforms are they have become an indispensable source of information during real-time events be it a natural disaster, terrorist attacks, political campaigns, epidemics or any other breaking news. Social media users are far ahead in posting up-to-date information than any news channels or news websites. As a result, many people rely on these platforms for getting real-time information in times of crisis and otherwise. Given the humungous data generated by social media users, it is challenging and at the same time imperative to analyze the trustworthiness and the veracity of the information. Although, a rich and accessible source of information, it is not a reliable one. On one hand the accessibility allows greater dissemination of important information, but on the other hand it also allows the spread of unauthentic information.

Indeed, trust and veracity are one of the major caveats of social media posts. To add to it, sometimes this happens that the user is posting some information and doesn't know that what he is posting is wrong. The restlessness of people to post latest information which in turn leads to greater likes, shares, comments and shares on their updates overpowers their ability to pause, think and research if what they are posting is even true or not or even if their information is coming from a credible source. Credible sources, on the other hand, like news websites, news channels' websites, blogs of reporters and journalists are themselves not completely trustworthy, to say nothing of others. They are always on verge of competition and trying to be the first one to bring the latest updates that they do not verify the veracity status of their information. Fake news generation is also a major issue with news websites. Although, some fake news are honest mistakes, but usually it is too late by the time the mistake is realized, the harm has already been done. Public figures like politicians, celebrities suffer the most due to these fake news and untrustworthy information sometimes forcing them to take extreme measures. Disasters are another crucial time when the credibility of information is most crucial. Although there are many people who genuinely want to help and post true or eye witness updates, there are many who simply pass on the information but proper verification. Although, they might not personally aim to cause mischief, but most people get trapped by the rumor mongers. For instance, during an earthquake in Chile rumors spread through Twitter that a volcano became active and there was a tsunami warning in Valparaiso [22]. Later, these reports were found to be false.

Thus, if not monitored, social media can be a breeding ground of rumors. Rumors not only give out false information (which may be intended or unintended) but can jeopardize the lives of people in some situations, tarnish the images of public figures in some and cause general unrest among the masses in others. This has been a great concern lately and many researchers have probed into the area of rumor detection and veracity.

Kumar and Sangwan [23] define rumor as, "any piece of information put out in public without sufficient knowledge and/or evidence to support it thus putting a question on its authenticity. It may be true, false or unspecified and is generated intentionally (attention seeking, self-ambitions, finger-pointing someone, prank, to spread fear and hatred) or unintentionally (error). Further, these can be personal as well as professional." They further provide the classification of information into rumor and non-rumor as shown in Fig. 7.

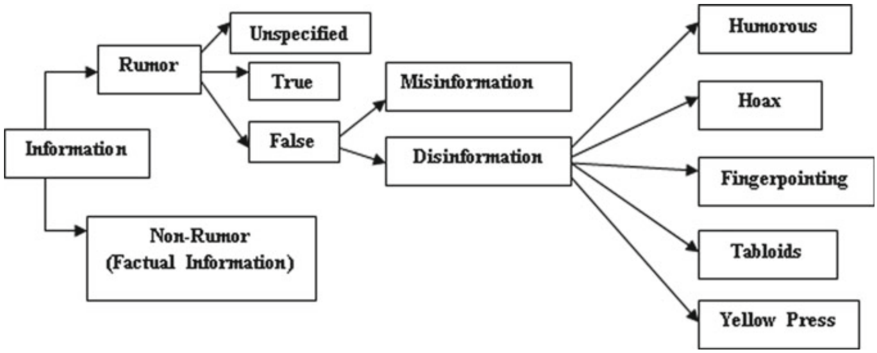

Fig. 7 Classification of rumors [23]

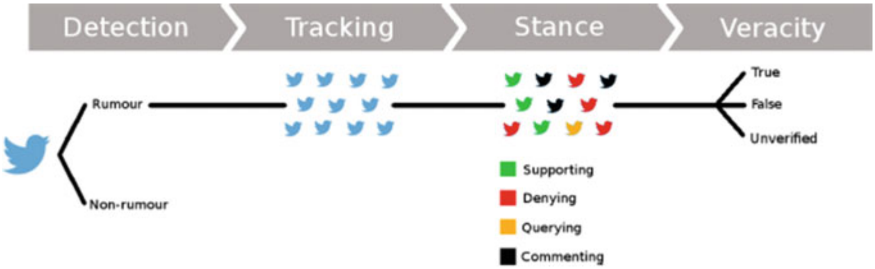

Fig. 8 Architecture of a rumor classification system [26]

There have been other attempts in classifying rumors such as based on veracity [24], credibility [25] and temporal characteristics [26].

Majority of the rumor detection work has been done on Twitter and Sina Weibo, a Chinese micro blogging platform with features similar to Twitter. Facebook, although one of the most popular social media on the web has not been explored much by researchers due to its restriction for data collection. Twitter provides its API for extracting data easily and free of cost, hence it is a viable choice for researchers. The data collections also depend largely on the type of rumors that one is gathering data for. It is easier for long standing rumors as one can search well defined hashtags and keywords for the rumors. Data collected from these rumors is useful for monitoring shift in the public opinion regarding the issue over a large span of time [26]. When it comes to gathering data for emerging rumors, it becomes more challenging due to the absence of well-defined keywords and hashtags. The rumor detection system can be used to detect these emerging rumors (Fig. 8).

Zubaiga et al. [26] propose the following components of a typical rumor classification system:

a. *Rumor Detection:* Given a stream of text/tweets as input this component can act as a binary classifier for detecting emerging rumors. PHEME and RumorEval are the two most popular publicly available annotated data sets for rumors.

b. *Rumor Tracking*: Given the identified rumors, this component keeps track of the posts discussing the rumor and keeping only those which are relevant and discarding the irrelevant ones. This component of the rumor detection system comes into play after a rumor has been detected and the subsequent posts regarding the rumor are tracked. The posts can either be filtered using relevant keywords that are monitored thus restricting the scope of the rumor tracking system or it can be input a stream of posts that are not limited by the filtering keywords thus broadening the scope. The output of the rumor tracking component would be posts labeled as being 'related' or 'unrelated' to the rumor being tracked. The dataset of 10,000 tweets developed by is the most widely used for rumor tracking.

c. *Stance Classification*: Given the posts discussing a rumor as tracked by the previous component, stance classifier can act as a multi class classifier labeling each post into classes such as denying, supporting or questioning the rumor. Thus, it can be used to determine the orientation of each post in response to a source tweet. However, it can be omitted where the stance of the public is not considered useful, e.g., cases solely relying on input from experts or validation from authoritative sources. PHEME is a widely used dataset that is publicly available for stance classification and gives annotations for stance of each tweet as support, deny, query, comment.

d. *Veracity Classification*: This final component uses the labeled posts from the stance classifier and additionally data from other sources like news websites and, or other databases to determine a truth value for the rumor. The rumor can be verified to be true, debunked as being false or labeled as being unresolved yet. The performance of this component can be measured using measures such as accuracy, precision and recall. The RumorEval 2017 dataset under the SemEVal2017 task consist of 300 rumors with veracity labeled as being true, false or unverified.

The first work that focused on detecting emerging rumours was done by Zhao et al. [27]. They assume that the rumors will invoke questioning tweets inquiring about their veracity. This can be used to detect the tweets as being rumors. They have used SVM, and decision trees for their approach. Zubiaga et al. [28], on the other hand, proposed an approach that is able to learn context during a breaking news story and can determine whether a tweet is rumorous or not. They used the hypothesis that a single tweet might not be sufficient for knowing if its underlying story is a rumour, as there is insufficient context. They used Conditional Random Fields (CRF) as a sequential classifier that was able to learn the dynamics of reporting during an event, such that the classifier can decide, for every incoming tweet, if it is a rumor on the basis of what has been seen by the system up to that point regarding the event. Their method improved the result in comparison to the baseline classifier by Zhao et al. [27], The classifier achieved 0.667 in precision and 0.556 in the recall, compared to 0.410 and 0.065 respectively for the classifier by Zhao et al. [27].

Rumor tracking has been explored little with the early work by Qazvinian et al. [29] being the only prominent one. They performed automated rumor tracking on a collection 10,000 tweets and used supervised machine leaning on various categories of features like content, network and Twitter specific. They have also studied stance classification in their work However, the most pioneering work in stance classification has been done by Mendoza et al. [30] from which they found that Twitter users largely support true rumors. Their study indicates that stances to rumorous tweets can be helpful in predicting veracity. Veracity classification has been one of the most studied aspects of rumors. The research for this task was initiated by Castillo et al. [31] who studied credibility perceptions of rumors and is considered a baseline for many subsequent works regarding veracity. Kwon et al. [32] suggested a novel feature types: temporal, structural and linguistic and used random forest, SVM and logistic regression and perform feature selection with the latter giving the best results. In [33] they have analyzed feature stability and found that structural and temporal features identify rumors over a long-term window. However, these are available only at later stages of rumor spread.

Little work has been on rumor tracking and origin detection. These areas can be explored extensively. Also, a finer grain classification of rumors into misinformation, hoaxes etc. as classified by Kumar and Sangwan can be explored.

5.2 Sentiment Analysis and Opinion Mining for Social Well-Being

Sentiment Analysis and Opinion Mining is “the field of study that examines people’s opinions, sentiments, evaluations, attitudes, and emotions from written language” [34]. Due to the abundant volume of opinion rich Web data, for example, micro blogging, news and website reviews, etc. accessible online via Internet, a significant part of the recent research is concentrating on the ongoing area of text mining (TM) field which is called as Sentiment Analysis (SA).

The term “opinion mining” was first observed in 2003 in a research paper by Dave et al. by 2006 [35] and thereafter, it spread to include Web applications and various other social wellbeing domains like marketing for product and service reviews, entertainment, politics, business, recommendations etc. which in turn broaden its application spectra. With the rapid rise of peoples’ participation and communication over social web where they express their ideas, opinions or sentiments over a public forum; SA extracts the public mindset and the findings are being used in various domains [36]. Next section briefs about the “What, Why and How Factors-WWH” of SA for Social well-being.

5.2.1 Sentiment Analysis: What?

SA is computational identification and classification of sentiments which are being expressed as written text on any particular topic, blogs, product reviews etc. [37] and has emerged as a dynamic area of research. Individuals are expected to build up a system that can identify, distinguish, arrange and classify sentiments accurately.

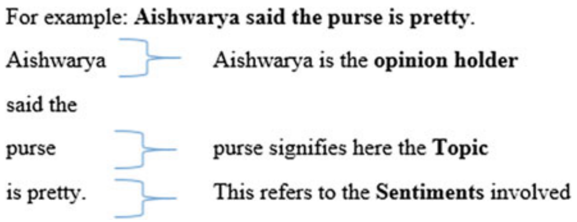
For example: Aishwarya said the purse is pretty.
Aishwarya
Aishwarya is the opinion holder
said the
purse
purse signifies here the Topic
is pretty.
This refers to the Sentiments involved

Fig. 9 Structure of a sentiment

The system should derive the mood and polarity of the sentiments [38]. SA analyzed the vast amount of heterogeneous information generated by social media users every day.

The ideology of SA is to search for opinions, identify the sentiments involved in it and classifying it based on its polarity as illustrated in Fig. 9.

5.2.2 Sentiment Analysis: Why?

SA is an emerging area of research in Natural Language Processing (NLP) and is widely studied in the fields of data, text and web mining. SA is not only limited to the field of computer science, but it also has applications in management and social sciences. The growing significance of SA accords with the development of social media (SM) such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks etc. [39]. This is the first time when a large volume of digitized opinionated data is available to us ready to be analyzed. Today, people express opinions largely through social media thus increasing the reliance on social networking sites Facebook, Twitter, etc. for decision making which extensively marks the increasing significance of SA in our daily lives [40].

For example, if an individual need to visit a place, instead of asking his friends, relatives or agents, etc., he simply turns on to its online real-time reviews of the visitors before taking any decision. Also, in terms of business management, if a client needs to buy a product, he first reads all its reviews and then eventually reaches to a decision whether to buy or not buy that product. Hence, we see that this content is available in huge quantity and is too vast to be analyzed meticulously, effectively and efficiently. Here comes the role of SA, which classifies the opinions based on its polarities.

5.2.3 Sentiment Analysis: How?

Sentiment analysis (SA) systems are being applied for practically all the social well-beings i.e. to almost all the business and social domain as opinions are central to most human activities and are considered to be the key influencers of our behaviors as well. Our beliefs and perceptions of reality, and therefore the selections we tend to make are basically conditioned on however others see and assess the world. For this reason, when we need to make a decision we regularly hunt down the opinions of others. This can be true not just for people, however conjointly for organizations as well. Various Web 2.0 tools and technologies can be used for efficiently evaluating the user opinions in real time [41].

5.2.4 Applications of Sentiment Analysis for Social Well-Being

With the rapid growth of SM, SA has become one of the most upcoming research areas in NLP. Its application is also widespread, ranging from business services, health cares, commercialization and industrialization to political campaigns [42–48] etc. It also involves various types of social well-being such as:

- ***Tracking collective user opinions for rating of the products and services (E-commerce and Product analytics)***
- ***Analyzing consumer trends, competitors and market buzz (E-industry)***
- ***Measuring response to company-related events and incidents (SM Monitoring)***
- ***Monitoring critical issues to prevent negative viral effects (Market Research and Analysis)***
- ***Customer Support (Voice of Customer)***
- ***Evaluating feedback in multiple languages***
- ***Brand Monitoring***

Thereafter, it is quite evident to say that sentiment analysis in the business can prove a major breakthrough for the complete brand revitalization. The key to running a successful business with sentiment data is the ability to exploit the unstructured data [49–53] for following actionable insights.

1. Business intelligence build up

Having insight and knowledge about the information at hand, eradicates the guess work enables to take and execute decisions in a timely fashion. Given the data today abundant with the sentiments about the established and the new products, estimating the customer retention rate has become much easier. Moreover, the reviews formed by sentiment analysis can prove helpful in adjusting the business strategy in accordance with the current market situation and hence lead to greater customer satisfaction. Keeping up and maintaining dynamicity is imperative in business intelligence and with the opinioned data at hand, one gets liberty.

2. Competitive advantage

Knowledge of the sentiments regarding the data of one's business competitors facilitates us to better our performance learning from their strengths and limitations. Sentiment Analysis and opinion mining can be also proved to be very helpful in foreseeing customer trends and forming business and marketing strategies accordingly.

3. Enhancing the customer experience

Customer satisfaction is the fertile ground on which any business flourishes. The satisfaction level of customers can be analyzed using sentiment analysis regarding various aspects of the business. This lets us know what has been well received in relation to products, services and customer support and what needs to be improved.

4. Brand brisking

The transition a business from being a startup to being a well-known and reputed brand depends on how well is the online marketing, social campaigning and content marketing done as well as how good is the customer support service. Sentiment analysis can help a business to know how its marketing strategies are being received, what needs to change and what needs to be improved.

The applications of sentiment analysis in E-commerce, product analysis, E-industry, market research and analysis [54–60] are overwhelming. It is also very useful in checking social media as it lets us get an indication of the broader public view on various topics. The applications of which are broad and powerful.

Few of the examples are as follows:

- The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential elections. Being able to quickly see the sentiment behind everything from forum posts to news articles means being better able to strategize and plan for future [61].
- Expedia Canada took advantage of this technique when they noticed there was a steady increase in negative feedback to the music used in one of their television advertisements [62]. Sentiment analysis conducted by the brand revealed that the music played on the commercial had become irritating after multiple airings. Rather than chalking it up as a failure, Expedia was able to address the negative sentiments in a playful and self-knowing way by airing a new version of the advertisement which featured the offending violin being smashed.

These days, sentiment analysis is used to address social problems as well, such as:

- Fighting infectious diseases

Understanding and influencing the confidence level and trust of the general public during an outbreak can help public health officials slow the spread of the disease.

- Preparing people for the future of work

How positive of negative a population feeling about the threat of automation is an indication of how well prepared they are for the upcoming changes.

- Promoting social inclusion

Efforts around the world to combat the spread of anti-immigrant, racist or anti-LGBTQ sentiment among young people increasingly rely on effective monitoring of the social media.

- Saving the environment

Whether raising public awareness about the impact of climate change or the pollution of the world's ocean, influencing public perceptions of environmental destruction is critical to changing course.

These applications of sentiment analysis have a very positive impact on the society and the environment, thus proving its huge relevance for social well-being.

5.3 Detection of Illicit Behavior and Bullying in Social Media

Web 2.0 is extending and evolving in terms of the volume, velocity and variety of information accessible online across various social media portals which affirms that the social media (SM) has global reach and has become widespread [63]. There are various benefits of SM but few people use it in the wrong way. One such illicit use of SM is to bully someone is known as Cyberbullying (CB). The term 'Cyberbullying' was coined by anti-bullying activist Bill Belsey in the year 2003 [64]. Tokunaga defined CB as "*any behavior performed through electronic or digital media by individuals or groups that repeatedly communicates hostile or aggressive messages intended to inflict harm or discomfort on others*" [65]. Different characteristics are highlighted by this definition like, the technology part, the antagonistic behavior of the act, reason to cause suffering, considered by most scholars to be crucial to the definition, and repetitiveness [63]. Cyberbullying or illicit behavior over social networks has already been designated as a major threat to public health by CDC (Centers for Disease Control and Prevention) [63]. CB can be possible through any of the media like mobile phones or using internet. CB may be done through emails, instant messages, chat rooms, blogs, images, video clip, text messages etc. [66, 67].

CB can be classified into direct CB and indirect CB. Direct CB is to harass or humiliate a person directly either through email, SMS etc. Indirect CB is done by posting harmful of humiliating content related to someone on SM. There are various types of CB explained in Table 4.

Table 4 Types of Cyberbullying

Types of CB	Description
Flaming	Online fight between people posting messages that contain offensive language [68]
Harassment	Posting messages in order to insult or threat the victim [68]
Denigration	Spreading rumors in order to harm the reputation of the victim [68]
Impersonation	Impersonate the victim and using this fake identity in damaging ways [68]
Outing and trickery	Acquiring the trust of the victim and the violating it by disclosing secrets of the victim [68]
Exclusion	Excluding the victim from groups or other online activities [68]

5.3.1 Incidences and Effects of Cyberbullying

Cyberbullying can happen at any age, sex and at any geographical location via any SM like Twitter, Facebook, social forums. It can be related to personal, racial, religious or cultural. The first reported case of CB was the case of Ryan Halligan of Vermont in 2003 [69]. Ryan was an American student who was constantly being bullied in person and via online by his classmates, that eventually forced him to commit suicide at the age of 13 only. But no legal actions could be taken against the culprits due to non-availability of proper laws pertaining to CB at that time. According to a survey, two third of the students in grade 7–9 belonging to middle class family or diverse communities in Calgary suffered from CB [70]. Another study [71] on a similar pattern exhibited that 15% of the students of 7th grade bullied others. Amongst all, fewer than 35% of the incidents were reported to the adults. Numerous studies showed that in school CB may occur due to lower academic performances, lower level of attachment and commitment to school or lower positivity in the school climate [71]. Age and demographic features also add to it. All this eventually affects the mind and body of the child and may lead to some of the disastrous effects such as committing suicides. CB may also occur due to the feelings of low self-confidence, suicidal ideation, annoyance, frustration, and various other emotional and psychological problems [72].

Research shows that persons who are facing CB are the ones who are victims of traditional bullying as well, but what makes CB more pervasive in the victim's life is the fact that CB can be reached at any given time of the day while in the traditional bullying, the bullying behavior usually happens during school time and stops once victims return home [64]. Therefore, the persistence of the cyber bullying behaviors may result in even stronger negative outcomes than traditional bullying [65, 73]. Moreover, CB has all those risk factors that are there in traditional bullying.

In addition, it has other aspects also that are very dangerous and can't be ignored like sometimes because of ignorance about the drawbacks and risks of sharing personal information over the Internet or sharing passwords on the internet, communicating with unknown people, a very little control exerted over personal information, which may lead towards CB [74]. Therefore, rather than being physically strong, CB tend to be more technologically savvy and should be able to catch the loopholes and remove them. It should work in a way so as to hide the electronic trails of victims, and use bullying "repertoire," which now includes identity theft, account hacking, infecting a victim's computer, impersonation, or posting embarrassing content [63].

In the year 2015, it was shown in a report in Italy that the users of 11–17 ages are bullied through email, SMS, or social networks. In the year 2016, there were 230 cases of cyberbullying that had been reported [75]. Recently, the Anti-cyberbullying Law, Law No. 71/17, officially came into force after getting approved by the parliament of Italy, which targets to tackle online bullying of children after several high-profile cases in which victims have committed suicide [76].

5.3.2 Cyberbullying Detection

Cyberbullying leaves an impression on the minds of cyber-victim. Thus, it is the need of the hour to detect CB. The process of CB detection is divided into different phases. It involves pre-processing of data, feature extraction, feature selection and classification of messages in bullying and non-bullying as shown in Fig. 10.

The first phase involves pre-processing of data, such as cleaning of data, correcting words, handling missing values, etc. Features are extracted from the pre-processed data, such as n-grams, skip-grams, profane words etc. Features are selected from

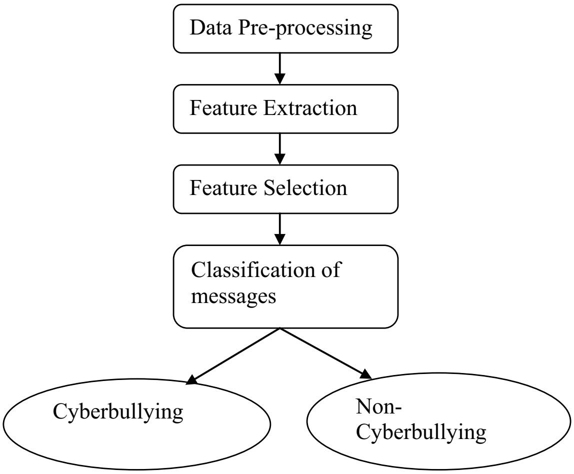

Fig. 10 Generic CB classification process

extracted features which are best and most effective. Finally, classification of messages is done to classify messages into bullying and non-bullying.

There are few software available to deal with cyberbullying e.g., [77–79]. Though, filters generally work with a basic key word search and sentiments of the text cannot be understood by the filters. There are some filters that block such webpage which contains some offensive keyword, while there are some such filters that just shows a blank webpage if any bullying keyword is detected. However, there are various techniques by which the filters can be dodged very easily, and thus it is true that filters are not that efficient and effective method for cyberbullying detection. Using chat rooms, mobile communication and peer-to-peer networking the blocked content can bypass central servers that maintain the filters. Another limitation is that the filtering methods have to be set up and maintained manually.

Various cyber bullying detection techniques has been applied by the researchers in the past few years in order to classify the bully comments more accurately and to make such models that will work quietly efficiently for the required task.

The Table below shows a few examples of the recent work done by various researchers for detection of cyber bullying. The Table 5 contains columns that show, which dataset used by the researchers to work on, what type of techniques and tools that have used, which type of social media site they have referred and also what results they have got.

There are several deep learning methods also that nowadays are being used in order to detect this illicit behavior over social networking sites. Deep learning methods, like CNN, RNN, are more efficient and has been proved more accurate by the researchers.

6 Conclusion

The chapter presents various aspects of Big Data including its characteristics, types, applications, etc. with the special attention on multimedia social big data which constitutes of data generated on social media sites and that is progressively growing. The data generated online not only consist of text, but images, videos, sensor information, emails, etc., which makes it complex in nature and difficult to store. There are some tools specifically used for storage and analysis of big data. Generation of big data cannot be stopped, but if utilized properly, it can be used in various research domains. Researchers are working on many applications of big data which includes rumor detection, sentiment analysis, sarcasm detection, cyber bullying detection, which have been explained in the chapter. As the data increases, it puts pressure on internet, web, applications, technology, etc. and demands advanced technology tools and software for processing of big data which could be challenging in the future. Some other applications of big data include fraud detection (credit card fraud, online auction fraud, insurance fraud, etc.), analysis of sales trends based on customer's buying history, applications in healthcare (personalized medicine and prescriptive analytics, effective treatment, drug side effects), applications in manufacturing (product

Table 5 State-of-art of cyber bullying

S.no.	Author	Year	Techniques	Social media	Data set	Tools	Results
1	Agrawal and Awekar	2018	Deep neural network	Twitter	Posts from Formspring, Twitter, Wikipedia	CNN, LSTM, BLSTM, and BLSTM with attention	DNN based models coupled with transfer learning beat the best-known results for all three datasets
2	Chen et al.	2018	CNN, Word2vec, Glove	Twitter	Comments from Twitter	Wordnet, porter	The proposed CNN model with 2-dimensional TF-IDF matrix results in improvement (with accuracy equals to 0.92and Macro-AUC equals to 0.98) compared with the baseline SVM and logistic regression methods
3	Andriansyah et al.	2010	SVM	Instagram	Comments on Instagram accounts of some Indonesian celebrities	R language, R Studio ide	The SVM model is able to classify the test set with an accuracy of 79.412%
4	Ting et al.	2017	Social network analysis, data mining	Facebook, Twitter, Ptt and CK101	100 posts from each website like Facebook, Twitter, ptt and ck101	–	The precision accuracy is around **0.79** and the recall is **0.71**
5	Lorenz and Kikkas	2013			Data was collected by means of closed questions or Likert scale options. (study was performed twice 2009, 2012)		There were only 1–3 phones in every class that were useful as a learning device

(continued)

Table 5 (continued)

S.no.	Author	Year	Techniques	Social media	Data set	Tools	Results
6	Badri et al.	2016	T-tests, chi square		More than 31,000 children from private and public schools participated in the online survey. More than 31,000 children from private and public schools participated in the online survey.		A high home access to the Internet of 91.7%. There is a negative correlation between time spent on social networks and perceived student performance in specific subjects
7	Bin et al.	2018	Random forest classifier	Reddit	Kaggle, It consisted of 6,594 raw comments	Skip-gram model architecture, Gensism (python library), natural language tool-kit	This model outperforms both the pre-trained word vectors, as well as the handcrafted methods by 3% more in AUC and 12% more on precision
8	Alduailej and khan	2017	Text mining, lexical approach	Twitter	Arabic tweet conversation from Twitter	RapidMiner, Python	Using text mining techniques cyberbullying can be detected automatically. The creation of lexicon of offensive Arabic words is the second approach that is presented in the paper

quality, supply planning, defects tracking), cybersecurity and intelligence, weather forecasting, traffic optimization, and many more.

References

1. J. Oliverio, A survey of social media, big data, data mining, and analytics. J. Ind. Integr. Manag. 1850003 (2018)
2. D. Borth, T. Chen, R. Ji, S.-F. Chang, SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content, in *Proceedings of the 21st ACM international conference on Multimedia, 21–25 October 2013* (Barcelona, Spain, 2013), https://doi.org/10.1145/2502081.2502268
3. A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, *03–06 December, 2012* (Lake Tahoe, Nevada, 2012), pp. 1097–1105
4. J. Weston, S. Bengio, N. Usunier, Wsabie: scaling up to large vocabulary image annotation, in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, *16–22 July 2011* (Barcelona, Catalonia, Spain, 2011), pp. 2764–2770, https://doi.org/10.5591/978-1-57735-516-8/ijcai11-460
5. M. Wang, D. Cao, L. Li, S. Li, R. Ji, Microblog sentiment analysis based on cross-media bag-of-words model, in *Proceedings of International Conference on Internet Multimedia Computing and Service, 10–12 July 2014* (Xiamen, China, 2014), https://doi.org/10.1145/2632856.2632912
6. A.B. Alencar, M.C.F. de Oliveira, F.V. Paulovich, Seeing beyond reading: a survey on visual text analytics. Wiley Interdiscip. Rev. Data Min. Knowl. Discov.**2**(6), 476–492 (2012)
7. I.E. Fisher, et al., The role of text analytics and information retrieval in the accounting domain. J. Emerg. Technol. Account. **7**(1), 1–24 (2010)
8. X. Hu, H. Liu, Text analytics in social media, in *Mining Text Data*, (Springer, Boston, MA, 2012), pp. 385–414
9. C.C. Aggarwal, H. Wang, Text mining in social networks, in *Social Network Data Analytics* (Springer, Boston, MA, 2011), pp. 353–378
10. Tobias Schreck, Daniel Keim, Visual analysis of social media data. Computer **46**(5), 68–75 (2013)
11. K. O'Halloran, A. Chua, A. Podlasov, The role of images in social media analytics: a multimodal digital humanities approach, in *Visual Communication* (De Gruyter, 2014), pp. 565–588
12. N. Diakopoulos, M. Naaman, F. Kivran-Swaine, Diamonds in the rough: social media visual analytics for journalistic inquiry. in *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)* (IEEE, 2010)
13. Bogdan Batrinca, Philip C. Treleaven, Social media analytics: a survey of techniques, tools and platforms. AI Soc. **30**(1), 89–116 (2015)
14. Tobias Schreck, Daniel Keim, Visual analysis of social media data. Computer **46**(5), 68–75 (2013)
15. W. Mason, J.W. Vaughan, H. Wallach, Mach. Learn. **95**, 257 (2014). https://doi.org/10.1007/s10994-013-5426-8
16. X. Wang, J. Yang, X. Teng et al., Feature selection based on rough sets and particle swarm optimization. Pattern Recogn. Lett. **28**(4), 459–471 (2007)
17. M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
18. Mohammad Ahmadi, Parthasarati Dileepan, K. Wheatley Kathleen, A SWOT analysis of big data. J. Educ. Bus. **91**, 1–6 (2016). https://doi.org/10.1080/08832323.2016.1181045
19. R. Talib, M.K. Hanif, S. Ayesha, F. Fatima, Text mining: techniques, applications and issues. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **7**(11) (2016)

20. P. Vashisht, V. Gupta, (2015). Big data analytics techniques: a survey, pp. 264–269. https://doi.org/10.1109/icgciot.2015.7380470
21. R. Reka Dr, K. Saraswathi, K. Sujatha Dr, A review on big data analytics. Asian J. Appl. Sci. Technol. (AJAST) **1**(1), 233–234 (2017)
22. Carlos Castillo, Marcelo Mendoza, Barbara Poblete, Predicting information credibility in time-sensitive social media. Internet Res. **23**(5), 560–588 (2013)
23. A. Kumar, S.R. Sangwan, Rumour detection using machine learning techniques on social media, in *International Conference on Innovative Computing and Communication*. Lecture Notes in Networks and Systems (Springer, 2018)
24. A. Zubiaga, M. Liakata, R. Procter, G.W.S. Hoi, P. Tolmie, Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS One **11**(3), 1–29 (2016)
25. M.E. Jaeger, S. Anthony, R.L. Rosnow, Who hears what from whom and with what effect a study of rumor. Personal. Soc. Psychol. Bull. **6**(3), 473–478 (1980)
26. A. Zubiaga, et al., Detection and resolution of rumours in social media: a survey. ACM Comput. Surv. (CSUR) **51**(2), 32 (2018)
27. Z. Zhao, P. Resnick, Q. Mei, Enquiring minds: early detection of rumors in social media from enquiry posts, in *Proceedings of the 24th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee, 2015)
28. A. Zubiaga, M. Liakata, R. Procter, Learning reporting dynamics during breaking news for rumour detection in social media (2016). arXiv:1610.07363
29. V. Qazvinian, et al., Rumor has it: identifying misinformation in microblogs, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2011)
30. M. Mendoza, B. Poblete, C. Castillo, Twitter under crisis: can we trust what we RT? in *Proceedings of the first workshop on social media analytics* (ACM, 2010)
31. C. Castillo, M. Mendoza, B. Poblete, Information credibility on Twitter, in *Proceedings of the 20th international conference on World wide web* (ACM, 2011)
32. S. Kwon, et al., Prominent features of rumor propagation in online social media, in *2013 IEEE 13th International Conference on Data Mining* (IEEE, 2013)
33. Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Rumor detection over varying time windows. PLoS One **12**(1), e0168344 (2017)
34. A. Kumar, T.M. Sebastian, Sentiment analysis on Twitter. IJCSI Int. J. Comput. Sci. **9**(4), 372–378 (2012)
35. K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in *Proceedings of the 12th international conference on World Wide Web* (ACM, 2003), pp. 519–528
36. A. Kumar, A. Sharma, A. Socio-sentic framework for sustainable agricultural governance. Sustain. Comput. Inform. Syst. (2018)
37. B. Pang, L. Lee, Opinion mining and sentiment analysis. Found. Trends Inf. Retr. J. **2**(2), 1–135 (2008)
38. A. Kumar, T. Sebastian, Sentiment analysis: A perspective on its past, present and future. Int. J. Intell. Syst. Appl. **10**, 1–14 (2012)
39. A. Kumar, A. Jaiswal, Empirical Study of Twitter and tumblr for sentiment analysis using soft computing techniques, in *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1 (2017)
40. B. Liu, *Sentiment Analysis Mining Opinions, Sentiments, and Emotions* (Cambridge University Press, Chicago, 2015)
41. A. Kumar, V. Dabas, A social media complaint workflow automation tool using sentiment intelligence, in *Proceedings of The World Congress on Engineering 2016*. Lecture Notes in Engineering and Computer Science (2016), pp. 176–181
42. A. Kumar, A. Joshi, Ontology Driven Sentiment Analysis on Social Web for Government Intelligence, in *Special Collection on eGovernment Innovation in India* (2017), pp. 134–139

43. E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**, 15–21 (2013)
44. R. Feldman, Techniques and applications for sentiment analysis. Commun. ACM **56**, 82–89 (2013)
45. A. Montoyo, P. Martínez-Barco, A. Balahur, An overview of the current state of the area and envisaged developments. Decis. Support Syst. **53**, 675–679 (2012)
46. S. Finn, E. Mustafaraj, Learning to discover political activism in the Twitter verse. KI-KünstlicheIntelligenz **27**, 17–24 (2013)
47. A. Trilla, F. Alias, Sentence-based sentiment analysis for expressive text-to-speech. IEEE Trans. Audio Speech Lang. Process. **21**, 223–233 (2013)
48. S. Tuarob, C.S. Tucker, M. Salathe, N. Ram, An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. J. Biomed. Inform. **49**, 255–268 (2014)
49. J. Brynielsson, F. Johansson, C. Jonsson, A. Westling, Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. Secur. Inform. **3**, 1–11 (2014)
50. P. Burnap, M.L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, A. Voss, Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. Soc. Netw. Anal. Min. **4**, 1–14 (2014)
51. A. Makazhanov, D. Rafiei, M. Waqar, Predicting political preference of Twitter users. Soc. Netw. Anal. Min. **4**, 1–15 (2014)
52. P. Bogdanov, M. Busch, J. Moehlis, A.K. Singh, B.K. Szymanski, Modeling individual topic-specific behavior and influence backbone networks in social media. Soc. Netw. Anal. Min. **4**, 1–16 (2014)
53. X. Fu, Y. Shen, Study of collective user behaviour in Twitter: a fuzzy approach. Neural Comput. Appl. **25**, 1603–1614 (2014)
54. X. Chen, M. Vorvoreanu, K. Madhavan, Mining social media data for understanding students' learning experiences. IEEE Trans. Learn. Technol. **7**, 246–259 (2014)
55. P. Burnap, M.L. Williams, Cyber hate speech on Twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet **7**, 223–242 (2015)
56. A. Zubiaga, D. Spina, R. Martinez, V. Fresno, Real-time classification of Twitter trends. J. Assoc. Inf. Sci. Technol. **66**, 462–473 (2015)
57. P. Andriotis, G. Oikonomou, T. Tryfonas, S. Li, Highlighting relationships of a smartphone's social ecosystem in potentially large investigations. IEEE Trans. Cybern. **46**, 1974–1985 (2016)
58. P. Burnap, M.L. Williams, Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci. **5**, 1–15 (2016)
59. N. Oliveira, P. Cortez, N. Areal, The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Syst. Appl. **73**, 125–144 (2017)
60. A. Singh, N. Shukla, N. Mishra, Social media data analytics to improve supply chain management in food industries. Transp. Res. Part E Logist. Transp. Rev. **114**, 398–415 (2018)
61. H. Wang, D. Can, A. Kazemzadeh, F. Bar, S. Narayanan, A system for real-time Twitter sentiment analysis of 2012 us presidential election cycle, in *Proceedings of the ACL 2012 System Demonstrations* (Association for Computational Linguistics, 2012), pp. 115–120
62. Understanding sentiment analysis: what it is & why it's used, https://www.brandwatch.com/blog/understanding-sentiment-analysis/. Accessed 19 Oct 2018
63. E. Aboujaoude, M.W. Savage, V. Starcevic, W.O. Salame, Cyberbullying: review of an old problem gone viral. J. Adolesc. Health **57**(1), 10–18 (2015). https://doi.org/10.1016/j.jadohealth.2015.04.011
64. M.A. Campbell, Cyber bullying: an old problem in a new guise? J. Psychol. Couns. Sch. **15**(1), 68–76 (2005)
65. Tokunaga Following you home from school, A critical review and synthesis of research on cyberbullying victimization. Comput. Hum. Behav. **26**, 277–287 (2010). https://doi.org/10.1016/j.chb.2009.11.014

66. Centers for Disease Control and Prevention. Youth violence: technology and youth protecting your child from electronic aggression (2014), http://www.cdc.gov/violenceprevention/pdf/ea-tipsheet-a.pdf. Accessed 11 Sept 2017
67. P.K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippett, Cyberbullying: its nature and impact in secondary school pupils. J. Child Psychol. Psychiatry **49**(4), 376–385 (2008). https://doi.org/10.1111/j.1469-7610.2007.01846
68. G. Sarna, M.P. Bhatia, Content based approach to find the credibility of user in social networks: an application of cyberbullying. Int. J. Mach. Learn. Cybernet. **8**(2), 677–689 (2017)
69. All you need to know about anti-bullying laws in India, https://blog.ipleaders.in/anti-bullying-laws/ Accessed 14 July 2018
70. Qing Li, Cyberbullying in high schools: a study of students' behaviors and beliefs about this new phenomenon. J. Aggress. Maltreatment Trauma **19**(4), 372–392 (2010). https://doi.org/10.1080/10926771003788979
71. Qing Li, Cyberbullying in high schools: a study of students' behaviors and beliefs about this new phenomenon. J. Aggress. Maltreatment Trauma **19**(4), 372–392 (2010). https://doi.org/10.1080/10926771003788979
72. J. Wang, T.R. Nansel, R.J. Iannotti, Cyber bullying and traditional bullying: differential association with depression. J. Adolesc. Health **48**(4), 415–417 (2011)
73. M.P. Hamm, A.S. Newton, A. Chisholm, J. Shulhan, A. Milne, P. Sundar et al., Prevalence and effect of cyberbullying on children and young people: a scoping review of social media studies. JAMA Pediatr. **169**(8), 770–777 (2015). https://doi.org/10.1001/jamapediatrics.2015.0944
74. J.A. Casas, R. Del Rey, R. Ortega-Ruiz, Bullying and cyberbullying: convergent and divergent predictor variables. Comput. Hum. Behav. **29**, 580–587 (2013). https://doi.org/10.1016/j.chb.2012.11.015
75. Commissariato di PS, Una vita da social, https://www.commissariatodips.it/uploads/media/Comunicato_stampa_Una_vita_da_social_4__edizione_2017.pdf. Accessed 28 Nov 2017
76. Law n. 71/17 of 29/05/2017, GU n. 127 of 03/06/2017. Senatodella Repubblica, http://www.senato.it/leg/17/BGT/Schede/Ddliter/43814.htm. Accessed 11 Sept 2017
77. Bsecure, http://www.safesearchkids.com/BSecure.html
78. Cyber Patrol, http://www.cyberpatrol.com/cpparentalcontrols.asp
79. eBlaster, http://www.eblaster.com/