Ali *et al. Journal of Big Data* (2022) 9:64
https://doi.org/10.1186/s40537-022-00620-4

Journal of Big Data

RESEARCH Open Access

Big social data as a service (BSDaaS): a service composition framework for social media analysis

Kashif Ali*, Margaret Hamilton, Charles Thevathayan and Xiuzhen Zhang

*Correspondence:
Kashif.ali@rmit.edu.au

School of Science, Royal Melbourne Institute of Technology University, 124 La Trobe Street, Melbourne, Australia

Abstract

Social media provides an infrastructure where users can share their data at an unprecedented speed without worrying about storage and processing. Social media data has grown exponentially and now there is major interest in extracting any useful information from the social media data to apply in various domains. Currently, there are various tools available to analyze the large amounts of social media data. However, these tools do not consider the diversity of the social media data, and treat social media as a uniform data source with similar features. Thus, these tools lack the flexibility to dynamically process and analyze the social media data according to its diverse features. In this paper, we develop a 'Big Social Data as a Service' (BSDaaS) composition framework that extracts the data from various social media platforms, and transforms it into useful information. The framework provides a quality model to capture the dynamic features of social media data. In addition, our framework dynamically assesses the quality features of the social media data and composes appropriate services required for various information analyses. We present a social media based sentiment analysis system as a motivating scenario and conduct experiments using real-world datasets to show the efficiency of our approach.

Keywords: Big data analysis, Service orientation, Social information services, Sentiment analysis, Service composition, Service quality

Introduction

The term big data is used to describe the large number of complex datasets that are beyond the ability of manual techniques and tools to collect, manage, process and analyze within reasonable time-frames [1]. Although big data is not a new concept or idea, there is no clear definition for 'Big Data'. Thus, it is defined on the basis of its characteristics. Initially big data was defined by its three characteristics, i.e., 3Vs, however later more characteristics and dimensions are added to the definition of big data [2]. Consequently, the big data is defined based on the eight characteristics also known as 8Vs [3]:

- Volume: the size of the data in measurable units such as terabytes (TB), petabytes (PB), etc.

- Variety: the type of the data. For example, different sources will produce big data in various formats such mobile devices, wireless sensor networks, software logs and social media platforms.
- Velocity: the frequency or speed of the data generation. For instance, the amount of data generated every millisecond, second, minute, hour, day, etc.
- Veracity: the meaningfulness of the data. This property determines the certainty in quality of data.
- Value: the actual business or application worth of the data. For instance, it pertains that how much knowledge or decision making information can be extracted from the data.
- Variability: the fluctuation between the generation and consumption of the data. For example, on different events, the data streaming exceeds the data consumption by business applications. On the other hand, occasionally, the data production decreases while the need of data consumption remains same.
- Viscosity: the resistance to the data flow. For example, the consistency of the data flow coming from various sources can be impacted due to changing rules and needs.
- Virality: the speed of propagation of the data. For instance, it specifies that how quickly data is shared within the network.

As a part of big data family, social media data has retained a unique position. Social media platforms, i.e., social information services, such as Facebook, Twitter, Youtube, etc., have emerged as free sources of big data. Big data generated by social information services termed 'Big Social Data' contains rich information such as public opinion and sentiment of social media users, i.e., social sensors [4]. Social information services, benefiting from cheap Internet access, have become a driving force in online data generation and sharing. Similar to big data, the 8Vs properties can be applied to social information services as follows:

- Social information service volume: According to statista.com, by the end of 2018, there will be 2.44 billion social sensors sharing their data online. As a result, the increasing number of social sensors causes social information service data size to grow exponentially.
- Social information service variety: There are six types of social information services: collaborative projects, blogs and micro-blogs, social networking sites, content communities, virtual games and social worlds. These types are classified on the basis of their primary usage [5]. The data produced by these services is unstructured and available in various formats (e.g., audio, video, images, text).
- Social information service velocity: Social information services generate data at an unprecedented speed. For instance, 500 million tweets are posted per day on Twitter, 55 million photos are sent per day on Instagram and 100 hours of videos are uploaded per minute on Youtube [6]. In 2021, on Facebook 2.90 billion users were monthly active. Daily 1.91 billion people on average logged in to Facebook[1].

[1] https://zephoria.com/top-15-valuable-facebook-statistics/.

- Social information service veracity: The massive production of the data by is a key characteristic of social information services [7]. Although this ever increasing data holds valuable information, the quality of the data varies and subjective to multiple measures. In addition, abstracting the high quality data with minimum level of noise remains a challenge.
- Social information service value: Social information services generate valuable data which is highly regarded in the decision making applications for customers [8]. Consumers often utilize the latest available online information before taking monetary decisions.
- Social information service variability: Social information services provide a unique way to share the real-time events with online users [9]. Any unexpected events (e.g., natural disasters) attract a large crowd of users who start generating bursts or streams of data in a short period of time.
- Social information service viscosity: The various types of social information services generate data with diverging characteristics [10]. For instance, the consistency of data captured from video streaming social information services and micro-blogging services remains volatile.
- Social information service virality: Over the time, social information services have become the main source of information spreading [11]. It merely takes just a few hours before any forms of digital content (e.g., news, videos, memes) go viral within the online community and spread across the globe.

The success of social information services has encouraged practitioners and researchers to utilize social big data in various domains such as business and marketing, product design, natural disaster and emergency management, politics and health surveillance [12]. Although social information services (e.g., Twitter, Facebook and Youtube) are used for online data sharing, their underlying mechanisms and social sensor contribution may vary [13]. In our previous work, a classification model to demonstrate the diverse data features of social information services was developed [14]. The effects of different topics of interests such as politics, entertainment, and health, on data features for various social information services were investigated [10]. For example, Twitter is a micro-blogging site which allows its users to share the data by using 280 characters long messages, i.e., tweets. Meanwhile, Facebook is a social networking site which allows its users to make personalized profiles, and share text messages, images, videos, etc., with their friends, family, and/or with public community. In contrast, Youtube is a video streaming site where viewers post their comments in response to the uploaded videos. Hence, the data generated by these dissimilar services has diverse features (e.g., size, text length, noise).

Despite the presence of 'variety' in social information services, current analysis tools and approaches treat social information services as a single entity with similar data features. For instance, there are several commercial social information service analysis tools available online. Most tools do not differentiate data by social information service features, and utilize uniform techniques for noise filtering and information extraction. In addition, these tools broadly focus on general-purpose information

search and analysis. Many tools are specifically dedicated to a single social information service (e.g., Twitter). Consequently, end users have to use multiple tools in an ad-hoc manner to analyze various social information services [15]. Using multiple tools is time consuming and provides inconsistent views of social sensors data. On the other hand, there are various traditional 'data-oriented' approaches available for analyzing social information service data. However, these techniques require multiple manual and time extensive activities such as manual dataset labeling, algorithm training and validation for information analysis [16]. Thus, we argue that in order to effectively analyze multiple social information services, it is necessary to dynamically process and analyze these services specifically according to their diverse data features.

In order to cope with the above challenges, we propose a 'Big Social Data as a Service' (BSDaaS) composition framework that extracts big social data from various social information services, and transforms it into useful information. Our framework utilizes the service oriented architecture (SOA) as a blue print to develop the BSDaaS framework. The SOA enables the usage of available services for the extraction, processing and analysis of big social data. Our framework devises a unique quality model to present the diverse features of big social data. The novelty of the proposed approach is that each social information service is processed and analyzed based on its unique data features, unlike traditional approaches where all social information services are treated at the same level. Moreover, the proposed framework dynamically extracts and assesses the data features of social information service and, based on the extracted features, appropriate services (e.g., data pre-processing, information extraction) for each social information service are composed.

To best of our knowledge, our proposed work distinguishes from traditional approaches by being first to leverage the SOA principals for social information service analysis. In this paper, we present social information service based sentiment analysis as a motivating scenario. However, the proposed framework is not limited to sentiment analysis, and it can be applied to other applications such as data mining and text mining. The main contributions of our work are as follows:

- A service composition framework that exploits the data quality features of social information services to compose big social data for various analysis applications. Our framework includes a multi-pipeline service composition mechanism which simultaneously processes multiple social information services and composes the aggregated results.
- A new formal model to define composite and component services for big social data analysis. A service quality model that captures the dynamic features of social information services based on big social data quality features.
- A dynamic mechanism which extracts and assesses the quality features of social information services. A quality model driven service composition method is proposed based on graph-planning.

The rest of the paper is organized as follows. Section II defines the problem statement and develops a motivating scenario. Section III highlights the work related to similar efforts. Section IV describes the details of the solution approach. Section V provides

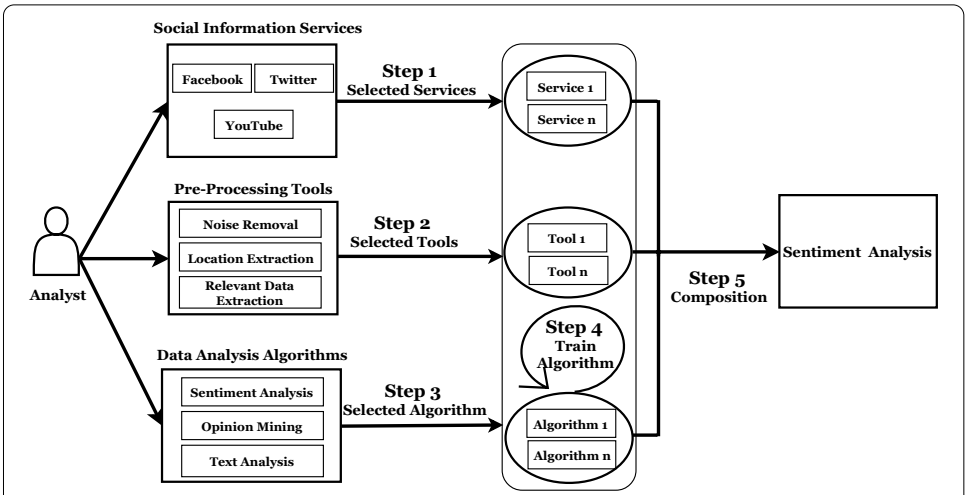


Fig. 1 Opinion analysis system design scenario

the details of the experimental implementation and evaluation results by using the real-world data. Section VI concludes the paper and provides some directions for future work.

Problem statement

In this section, first a motivating scenario is presented to illustrate the challenges that can occur when using a traditional social information services analysis approach. Then, the illustrated challenges in a set of problems are analyzed.

Motivating scenario

Let us assume that 'Bella' is a social information service analyst who is working for the 'National Interest First' political party. The party is gearing towards the upcoming presidential election. In order to define the vision for the election campaign, the party is interested in gathering public opinion and sentiment about presidential candidates, and on various political and policy issues. In addition, the party is interested to visualize the public opinion based on multiple constituencies, i.e., geographical locations.

In this scenario, Bella is given the task for developing an opinion analysis system to analyze the public opinion from social information services. Bella develops the opinion analysis system by using a typical sentiment analysis approach (see Fig. 1). She develops the system by combining several processes. She first selects various social information services as data sources for obtaining the public opinion. In particular, Bella selects only those social information services which have the geo-tagged data. Secondly, she defines a process model to remove the noise, i.e., irrelevant data, from the collected data. Thirdly, she uses a subset of the collected data to train a machine learning algorithm for various analysis tasks (e.g., sentiment classification, entity extraction). Finally, the analysis results are presented in various formats (e.g., charts, graphs) for comparison.

Problem analysis

Bella's approach has several limitations. To manage the above scenario, the sentiment analysis system problem is divided into following sub-problems.

- The geographical location information allow users to visualize the data based on social sensor locations [17]. However, many social information services do not provide the facility to geo-tag data. Bella would be required to use only social information services which provide the geo-tagged data. Consequently, she may intentionally miss several potential information sources.
- Social information services provide various functionalities (e.g., video streaming, image sharing) and offer different limitations (e.g., text length) for data sharing which effect the data quality. For instance, social sensors may intentionally use misspelled words and abbreviations. Moreover, social sensors may write text in an informal style which is heavily influenced by Internet language and may contain slang, emoticons and special characters [18]. Thus, the quality of data (e.g., noise level) varies in different services. Hence, one noise removal mechanism can not apply to all.
- The data collected from social information services contains subjective information such as opinion, sentiment (e.g., positive, negative, neutral) and emotion (e.g., happy, sad, angry) [19]. Furthermore, the topic of interests, i.e., data trends, on social information services keep changing. Thus, similar to noise removal, one specific analysis algorithm may not be applicable to all types of social information services. Consequently, Bella needs to constantly train and validate the algorithm(s) required for information extraction, which is time-consuming.

With the above limitations, the sentiment analysis system may not illustrate real public opinion. It requires constant involvement with laborious activities from the analyst (i.e., Bella) for algorithm training and validation.

In contrast, BSDaaS dynamically processes social information services based on their quality features. This enables the composition of appropriate services, e.g., noise removal, sentiment analysis, etc., based on relevant features for information extraction and integration.

Related work

Social information services have emerged as an integral part of big data family and provide the ability to develop new applications. Researchers and practitioners have explored the usage of social information services in diverse domains, for instance, in digital transformation and tourism [20]. This facilitates online communication between customers and decision makers. Social information services are also used in product design [21]. Business organizations have used social information services to identify requirements for developing new products. Social information services are also used in election campaigns to influence the decisions of the voters. US presidential candidates have successfully used Facebook and Twitter for their election campaign [22]. Social information services enable the surveillance and tracking

capabilities for epidemic monitoring. For example, blog posts are used to track the flu epidemic in social sensors [23]. Social information services can play a vital role in disaster and crisis events. Social information service data can be used to enhance emergency situation awareness and as early warning systems [24]. The efficacy of social information services in terrorism and violence events [25]. Governments are investing to track down terrorists with the help of social information services. Furthermore, despite the availability of traditional communication systems, social information services provide a platform for communication and relief activities [26].

Social information services are increasingly becoming a promising platform to unveil the social sensor activities based on time and space (i.e., geographical locations) [27]. Spatio-temporal information can be utilized to concentrate the efforts of government agencies on critical locations [28]. The effectiveness of spatio-temporal analysis of location based social information services data has been well demonstrated in a major natural disaster events [29]. Similarly, an extended framework is proposed to process the large spatio-temporal data from Twitter. As a case study, the Twitter data is analyzed for South East Queensland floods to demarcate the locations of disaster zones [30].

The data (e.g., text) shared on social information services is unstructured. One mechanism to extract the useful information from the textual data is sentiment analysis [31]. The main task of sentiment analysis is to classify a piece of text into three categories: positive, negative and neutral [32]. However, in broader term, sentiment analysis or opinion mining is a domain of Natural Language Processing (NLP) which extracts people's opinions, feelings, and emotions toward entities, individuals, topics, and events from written languages [33]. Sentiment analysis is a complex and multi-step process comprising of three sub-processes: data collection, data preprocessing (e.g., data cleansing, noise removal) and data analysis (i.e., information extraction) [34]. There are mainly three types of techniques for sentiment analysis: Machine learning based, Lexicon based and Hybrid [35]. The machine learning approaches are mainly based on statistical models to extract and classify sentiments; while the lexicon techniques are comprised of semantic dictionary or various corpus based approaches. Hybrid approaches combine both machine learning and lexicon driven techniques. Sentiment analysis has been used in various applications such as decision making, marketing, consumer feedback management, predicting polls and forecasting trading markets [36].

The service oriented architecture (SOA) provides the baseline framework for cloud services infrastructure. SOA has emerged as an architectural concept which promotes the practices to build applications by using services as smaller functioning block which are loosely coupled, reusable, agile, efficient and inter-operable [37]. SOA provides the power of abstraction which hides the development and implementation complexities and put focus on creating applications independent of hardware platforms, middleware, and programming languages [38]. One major benefit of SOA is that it enables the composition of new applications by reusing the existing services [39]. Services can be composed based on their functional properties. However, with the availability of multiple functionally similar services, one challenge arises: how to chose the most suitable service(s). Consequently, a number of approaches have emerged which utilizes the non-functional properties or quality of service (QoS) properties (e.g., price, response time) to find optimal services based on end user requirements [40].

With the increasing commercial usage of social information services and its data, there are various online tools which are used for monitoring, analytics and sentiment analysis. Some tools are dedicated to a specific and famous social information service. For instance, the tools such as LikeAlyzer[2] enable the analysis for Facebook public pages with multiple engagement metrics. Similarly, tools like Twitonomy[3] provide the visual analysis for Twitter based on different criteria (e.g., hastags, re-tweets) and display results such as trending tweets and issues, polarity and content ranking. There are also several tools available that support the simultaneous analysis of more than one social information service. For example, SocialMention[4], Hootsuite[5] and Klear[6] support analysis of more than three social information services at same time. Generally, these applications are used for concurrent engagement with online users, scheduling and publishing of contents such as promotional videos, images, and interactive messages. However, for the information extraction and analysis, these tools treat and analyze multiple social information services at the same level, and provide simple analyses dashboard setups such as social sensors clicks and likes, general sentiment, popular contents, hash-tags, etc., for end users. In addition, current tools lack the flexibility to compare the analysis results across social information services. Also, they lack the flexibility to compose information according to dynamic end user queries (e.g., spatio-temporal analysis).

The concept of 'big data' initially described as a large-scale data which can not be processed, analyzed and presented by using existing technologies and techniques [41]. Big data has no single source of data generation rather it is generated by multiple sources such as humans, sensors and business activities [42]. Cloud services provide an ideal infrastructure which is reliable, fault-tolerant, and scalable for big data collection, processing and analysis [43]. There are several research efforts [44] [45] [46] exploring the Big data in cloud services environment. However, these paradigms only support limited functionality and provide high-level views of big data at service level. Similarly, there are several efforts made for processing and analysing big social data [47] [48]. These techniques focus on data driven methods instead of dynamic processing and analysis of big social data by leveraging the service oriented architecture.

Solution approach: big social data as a service

The proposed solution approach is based on the service-oriented architecture (SOA) principles. By utilizing service orientation, the components or processes required for sentiment analysis are replaced with the available cloud resources (e.g., services). SOA provides several benefits over traditional sentiment analysis approaches including separation of concerns, dynamic composition and abstraction. Firstly, 'separation of concerns' enables the sentiment analysis system to determine a set of loosely coupled layers where each layer has a set of task specific services (e.g., APIs, Web Services). These services can be technology independent and provided by various service

[2] http://likealyzer.com/.

[3] https://www.twitonomy.com/.

[4] http://www.socialmention.com/.

[5] https://hootsuite.com/.

[6] https://klear.com/.

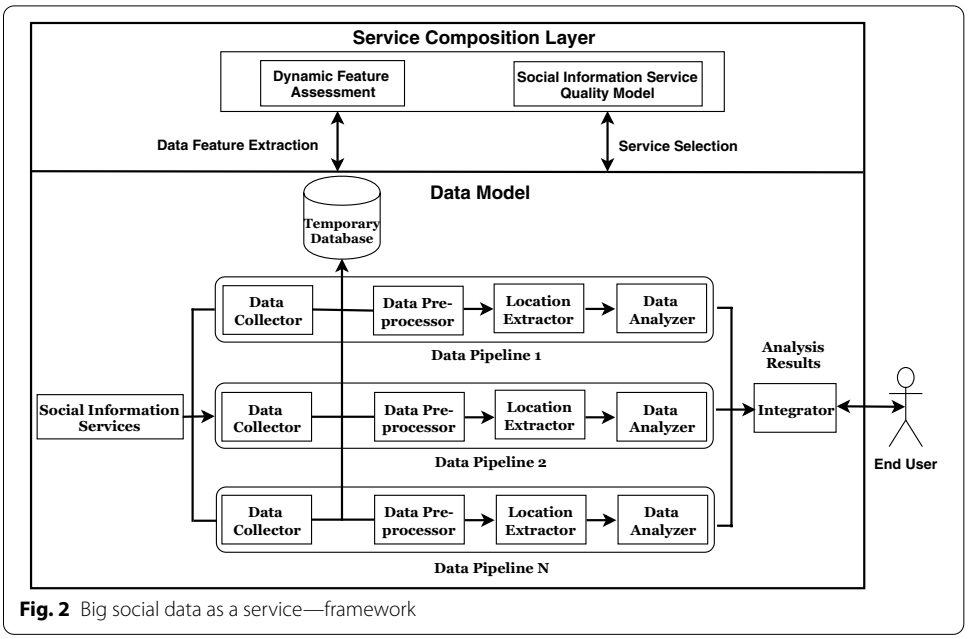


Fig. 2 Big social data as a service—framework

providers. Secondly, SOA provides the ability to dynamically select and compose services based on functional and QoS requirements. Finally, abstraction hides the implementation details from end users, and services are maintained by service providers. Thus, the time extensive tasks such as training and validations of service's underlying implementation do not involve the service users.

The solution approach (see Fig. 2) is divided into two parts: (1) Data Model (2) Service Composition Layer. The data model defines the data processing pipelines where each social information service is processed by a set of components and later results are aggregated. Each component is conceptualized as a set of functionally similar services. The data is shared between components via data messages, e.g., XML, CSV. The service composition layer extracts the data features of each social information service in a data pipeline and composes appropriate services.

In the following sub-sections, the details of the data model and the functionality of the service composition model is explained.

Data model

The data model defines a data pipeline architecture for big social data analysis. The data pipeline model consists of five data processing components (i.e., service layers): Data Collector, Data Pre-processor, Location Extractor, Data analyzer and Integrator. The data pipeline enables simultaneous processing of multiple social information services based on their data features. In particular, each social information service is processed by first four components, i.e., data pipeline. Initially the collected data is stored in a temporary database that enables the composition layer to select further services for the pipeline. Finally, the integrator composes results of each data pipeline, and presents data in various formats (e.g., tables, maps) as BSDaaS. Formal

definitions and details of the data processing components as services are provided as below:

Big social data as a service

We use a top-down approach to formally define the component services. The BSDaaS is defined as a composite service which is accessed by the end users.

Definition 1 The Big Social Data as a Service (BSDaaS) is a composite service which incorporates the required data processing services and provide analysis results to the end users. BSDaaS is a tuple of $< ID, AP, AF_i, Q_i >$, where

- *ID* is the service id.
- *AP* is the type of analysis application (e.g., sentiment analysis).
- *AFi* is set of functions (e.g., sentiment polarity) provided by application.
- *Qi* is a set $< q1, q2, ..., qn >$, where q_i denotes a QoS feature (e.g., price).

Data collector

The data collector gathers the data from various social information services. Many social information services provide a standardized mechanism, i.e., APIs, for searching and extracting the data. For example, Twitter provides search API to find the relevant data based keyword(s) (e.g., Trump, US President). In contrast, many social information services do not provide search APIs. Alternatively, Web scraping (i.e., HTML parsing) is used for the data gathering. The data collector saves the data into a temporary repository.

Definition 2 The data collector service *DCS* is a tuple of $< cid, k, ts-te, SN, q_i >$, where

- *cid* is the service id.
- *k* presents a set of keywords as an input to search the relevant data.
- *ts-te* present temporal bounds *ts*/*te* (start/end) time for data collection.
- *SN* is the targeted social information service.
- *qi* presents the set of QoS values (e.g., response time).

Data pre-processor

The data pre-processor removes the noise from the collected data. Each dataset is processed by following filtering order:

- Language filter: Due to the global nature of social information services, social sensors across the globe share their data written in various native languages (e.g., English). The language filter removes the data based on a given language.

- Extra data filter: The collected data often contains the unnecessary information (e.g., embedded URLs). For example, the tweet "Angry Ivanka Trump Walks Out Of Cosmo Interview youtu.be/nKxxlftcYyY via @YouTube" contains embedded URL. The extra data filter removes the unnecessary data. Also, the , repeated data items and data without time stamps is removed.
- Relevant data filter: The collected data often contains irrelevant data. However, the degree of relevant information can vary for each social information service. In order to retrieve the relevant data, there are several information retrieval techniques (e.g., relevance, ranking) [49] with different performance measures (e.g., throughput, quality of results). A probabilistic method is used to assess the relevance of each datasets, in order to apply appropriate data filtering technique.

Definition 3 The data pre-processing service *PRS* is a tuple of $< pid, fln, dp, qi >$, where

- *pid* is the service id.
- *fln* describes a set of noise removal filters.
- *dp* presents the relevant information extraction technique used in the service.
- *qi* defines the QoS properties (e.g., accuracy).

Location extractor

Many social information services do not provide the facility to social sensors for geo-tagging the data. Moreover, some social sensors intentionally choose not to expose their geographical information. Therefore, the location information is extracted by checking the mentions of location names (e.g., cities) in the text. The location extractor first parses the data, i.e., text, to identify locations. Secondly, it uses the geo-coding technique to assign geo-coordinates, i.e., longitude and latitude, for extracted locations. The data without location information is eliminated.

Definition 4 The location extractor service *LES* is defined as a tuple of $< lid, flc, qi >$, where

- *lid* is the service id.
- *flc* extracts locations from text and geo-tag the data by using a gazetteer.
- *qi* is a set of QoS properties (e.g., Throughput).

Data analyzer

The data analyzer consists of multiple data analysis services to extract the required information (e.g., sentiment, emotion). There are various commercial and non-commercial services available online which are specialized in analyzing the data (e.g., text)

of different types of social information services. The benefit of using online services is that these services are maintained and updated by service providers. Furthermore, these services also provide several non-functional features (e.g., price) to fulfill the QoS preferences.

Definition 5 The data analyzer service *DAS* is a tuple of $< aid, inf, dap, ln, qi >$, where

- *aid* is the service id.
- *inf* defines the subjective information (e.g., sentiment, entities) required from the data.
- *dap* defines the data features require for analysis service selection.
- *ln* presents the ability to perform the analysis on multiple languages.
- *qi* determines the set of QoS (e.g., precision).

Integrator

The integrator composes the results from multiple data pipelines based on the spatio-temporal requirements. Let us assume that M is a matrix of j number of data pipelines. A data pipeline P_j is comprised of four data processing components: data collection C_j, data preprocessing R_j, location extractor L_j, and data analyzer A_j. Each component has k number of candidate services. In matrix M, each tuple represents a data pipeline. A data pipeline P_j may be composed of multiple candidate services from each data processing component. Equation 2 presents the composition of data processing components for a data pipeline P_j of social information service *SN*.

$$M = \begin{Bmatrix} P_1 = \left(C1_{ci}^{ck}\right), \left(R1_{ri}^{rk}\right), \left(L1_{li}^{lk}\right), \left(A1_{ai}^{ak}\right) \\ P_2 = \left(C2_{ci}^{ck}\right), \left(R2_{ri}^{rk}\right), \left(L2_{li}^{lk}\right), \left(A2_{ai}^{ak}\right) \\ \cdots\cdots \\ P_j = \left(Cj_{ci}^{ck}\right), \left(Rj_{ri}^{rk}\right), \left(Lj_{li}^{lk}\right), \left(Aj_{ai}^{ak}\right) \end{Bmatrix} \tag{1}$$

$$P_j(SN_j) = \sum_{i=1}^{k} j.[C_{ci}^{ck} + R_{ri}^{rk} + L_{li}^{lk} + A_{ai}^{ak}] \tag{2}$$

It is possible that a data pipeline may provide results in various formats or data types. The integrator component incorporates the analysis results of all data pipelines into a cohesive format based on spatio-temporal features. Equation 3 shows the composition function $Com(S, T)$ for pipelines integration, S and T presents the spatio-temporal parameters.

$$Com(S, T) = \sum_{j=0}^{n} P_j(SN_j) \tag{3}$$

Definition 6 The data integrator service is a tuple of $< npl, space-time, df >$

- *npl* defines the set of data pipelines aggregated for final composition.
- *space*−*time* space and time clusters the large data into small segments based on location (e.g., states, cities) and time.
- *df* presents the final results in various formats (e.g., maps, table, charts).

Service composition layer

The service composition layer forms the backbone for the service selection and composition process. It comprises of two main components: Social Information Service Quality Model, Dynamic Feature Assessment. The quality model provides a set of features that enables the selection of appropriate services for a data pipeline composition. The dynamic feature assessment component abstracts the chunks of data of each social information service from the temporary database and assesses the quality based on the quality model. Finally, the composer selects the suitable services for data processing.

Social information service quality model

In this section, a quality model to capture various quality features is devised by using existing domains, i.e., sensor cloud computing, data quality assessment. The proposed QoS model is extensible. Currently following properties are used in the model:

Data volume: The data volume V_N determines the quantity of the collected data (e.g., number of tweets, comments) [50]. The data volume is useful to predict the amount of steps (e.g., time, space) required for the data processing. The data volume for social information service *SN* is calculated as following function:

$$f_{Volume}(SN) = V_N \tag{4}$$

Data richness: Social information services have rich data due to the contribution of multiple social sensors [51]. The high number of social sensors determines the confidence in final analysis. The richness R_{Sen} means the number of unique social sensors. It is used to select or reject a dataset based on satisfactory participation of social sensors. The data richness is calculated as following function:

$$f_{Richness}(SN) = R_{Sen} \tag{5}$$

Data freshness: The freshness implies that the data is recent and does not contain old data [50]. The freshness F_R determines the data data based on temporal properties. The freshness is calculated as below:

$$f_{Freshness}(SN) = F_R \mapsto \int_{ts}^{te} \Delta \tag{6}$$

where *ts* is the time, i.e., time stamp, of oldest data item, and *te* defines the time of latest data item in the dataset Δ.

Data relevance: Relevancy defines the extent to which the data provided by a social information service is applicable to a given topic [52]. The relevance property D_{Rel} implies the induction of an appropriate information retrieval mechanism to retain relevant data items. The relevancy of a dataset Δ of *SN* is calculated as follows:

$$D_{Rel}(SN_{\Delta}) = \frac{A \cap \Delta}{\Delta} \tag{7}$$

where A is number of relevant data items in a given dataset.

Text type: There are two types of online text writing styles: formal and informal [18]. In informal style T_{Inf}, the text is short and contains Internet language (e.g., abbreviations, slang). In formal style T_{Fo}, the text is written by using proper language skills. Information extraction from both types requires different text analysis techniques. Thus, the text type helps to determine best suited text analysis technique. The text type T_{Type} classifies a data item d_i into a text type as follows:

$$T_{Type}(SN_{\Delta}) = \forall d_i \in (\Delta) : \{Type \leftarrow T_{Fo} | T_{Inf}\} \tag{8}$$

Algorithm 1 Dynamic Feature Assessment

Input: Temporary Database DB, Social Information Services SN_i, Search Terms K, Stop Terms S
Output: Data Features List $F < SN_i(list) >$
1: Connect Connection (DB)
2: **for** (each $SN_i \mapsto DB$) **do**
3: Extract $SubSN_i(D) : SN_i \mapsto \Delta$
4: $F < SN_i > \leftarrow$ Get D_{Rel} (Δ, K, S)
5: $F < SN_i > \leftarrow$ Get T_{Type} (Δ)
6: Close Connection (DB)
7: return $F < SN_i(list) >$

Dynamic feature assessment

The high volume of big social data captured from multiple social information services poses challenge for sentiment analysis. The service composer has to analyze and estimate the quality features in a limited time, before the data processing services can be composed. Thus, it may not be possible to analyze a large volume of data in a limited time-frame. To cope with this problem, a probabilistic approach for quality feature assessment is proposed as follows:

Let us assume that $SN_i(D)$ represents a document in the temporary database DB which contains a finite set of comments $SN_i(D) = \{d1, d2, d3, ..., dn\}$ for a social information service SN. The dynamic feature assessment component extracts a random sample $SubSN_i(D)$, (i.e., subset) of 5% data for each $SN_i(D)$ in DB. Secondly, it analyzes and evaluates the two QoS features: Data Relevance and Text Type, by using two classifiers. Later, the service composer uses the extracted QoS features to compose available services. Algorithm 1 formally illustrates the details of the feature extraction and assessment process.

Data relevance classifier: To calculate the relevance of the data D_{Rel}, the classifier implements a simple boolean term frequency tf search model. Let us assume that $K = \{k1, k2, k3, ..., kn\}$ is a set of finite number of keywords used to query and obtain the data from a social information service SN_i. First, the keywords are used to semantically search each data item d_i in the $SubSN_i(D)$; d_i is considered a match if at least one keyword is matched. Thus, the data relevance D_{Rel} of $SubSN_i(D)$ is computed based on equation 7. Finally, this percentage of the relevant data items $P_{Match}(SubSN_i(D))$ in a given subset is determined as the probability value of D_{Rel}. Based on the required data

relevance threshold and data volume, the composer selects an appropriate information retrieval service to get the refined data. Generally, if a dataset has lower relevance probability, a probabilistic ranking model can be used, whereas, with the higher relevance, a simple retrieval model is implied.

Text type classifier: The text type classifier is a simple binary classifier which labels the data items of a subset $SubSN_i(D)$ into two textual categories: Formal and Informal. The binary classifier is trained with a subset of Internet slang (e.g., lol, omg) taken from Internet slang dictionary.[7] After successfully labeling all of the data items, the ratio of formal data to informal data is calculated. The classifier returns the higher ratio as the text type of the corresponding subset to the service composition layer.

Quality of service driven composition approach

In this section, a social information service quality driven service composition approach based on *Graph-Planning* is defined. Graph-Planning is an Artificial Intelligence (AI) based planning technique which uses states, propositions and relevant actions to solve a planning problem. The *Graph-Planning* technique has been used in dynamic workflow composition for data analysis [53].

In order to choose 'right' set of services for BSDaaS composition, the selection of required services is highly dependent on the quality of the social information service. Traditional QoS driven service composition techniques focus on the predefined user provided QoS constraints (e.g., throughput, accuracy) which are used as a service selection criteria. However, in BSDaaS, the QoS requirements are also determined on the quality of data gathered from social information services. Consequently, the composition process requires to proceed in conjunction data quality features. First, the services are selected based on data quality features. Secondly, it is possible that service composer matches more than one service. Thus, user provided QoS (e.g., throughput, accuracy) play a decisive role in the final service selection. There are several techniques [40] available that can be applied for user provided QoS based service selection. In this paper, the solution approach focuses on social information service quality driven service composition and assumes the existing techniques for performance QoS driven service selection. In [14], a semantic tagging model is established which tags the services in service repository with the quality model attributes (e.g., text type). The solution approach leverages the existing semantic model for service matching and retrieval in the proposed composition approach.

The proposed approach applies the service quality attributes as constraints for task planning. The proposed approach is defined in following three sections.

Composition model

The composition model takes the input of end users and formulates the service composition process as a multi-stage planning problem. The core concepts to form a *Graph-Plan* for service composition are explained as follows:

[7] https://www.internetslang.com/.

- Goal: The goal G_{BSDaaS} of planning problem consists of four sub-goals: G_{BSDaaS} $=\{G_{DCS}, G_{PRS}, G_{LES}, G_{DAS}\}$, where G_{DCS}, G_{PRS}, G_{LES} and G_{DAS} presents the sub-goals of data collection, data preprocessing, location extraction, and data analysis, respectively in a data pipeline composition.
- The planning problem: for data pipeline composition process is denoted as P. It is comprised of three elements: state transition Σ, initial state I_S and the goal G_{BSDaaS}, defined as $P = \{\Sigma, I_S, G_{BSDaaS}\}$.
- State transition: Σ consists of three sub-elements: set of states S, set of actions A and set of constraints C_s, defined as $\Sigma = \{S, A, C_s\}$.
- State: S_i consists of n number of tasks T_n, denoted as $S_i = \{T_1, T_2, .., T_n\}$. For example, data preprocessing consists of several noise removal tasks. For each task, there are m candidate services (i.e., noise removal filters) $T_i = \{ws_1, ws_2, ..., ws_m\}$ available.
- Action: A creates a respective task in the state transition based on constraints.
- Constraint: C_s is set of conditions adjacent to set of actions that should be true or false before an action creates a task. For instance, the after the completion of data collection G_{DCS} sub-goal, the data feature assessment should must be completed prior to the execution of data preprocessing G_{PRS} sub-goal.
- Task simulation: In *GraphPlan*, task simulation $F_n(T)$ simulates tasks based on conjunctive actions and constraints as $F_n(Ti) = \{A_i \mapsto C_{si}\}$. For example, an action *DataCollection* in G_{DCS} and respective constraint *Twitter* create the task of *TwitterServiceSelection* in the planning process.
- Trivial solution: Based on above concepts, each sub-goal is achieved by further decomposing the problem into states and corresponding tasks. A solution to P exists; if and only if all sub-goal states are intersected with a set of tasks which are reachable via initial task T_I; and final result set of the composition graph must not be empty: $G_{BSDaaS} \cap T_n^{>}(\{T_I\}) \neq \{NULL\}$.

Algorithm 2 Graph-Planning Algorithm

Input: Initial State I_S, Set of constraint C_s
Output: Composition plan G
Initialize (Init layer i, master table MT, graph G, init constraint C_{s0})
while (all constraints $S_i \rightarrow MT$) **do**
  expand graph(G, S_i)
  for (each S_i) **do**
    extract actions and constraints S_i:($A_i \mapsto C_{si}$)
    if valid state transition(ΣS_i) **then**
      add ($G \leftarrow$ Solution S)
    elseabort()
  $i = i + 1$
  if (i not changed) **then**
    abort()
return G

Graph-planning algorithm

The service composer in BSDaaS framework utilizes the composition model to generate the composition plan (i.e., composition graph). The proposed algorithm is extended based on *Automated Graph-Planning* technique [53]. In proposed method, the algorithm is limited to *Forward-Search* only to devise a *Graph-Plan*, in order to find a

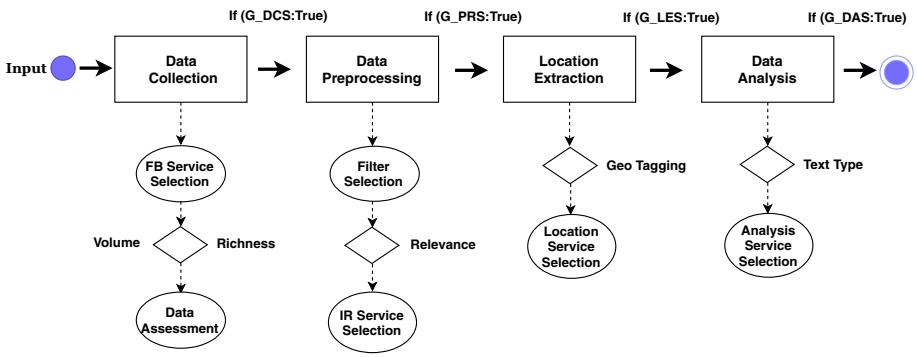


Fig. 3 Facebook data pipeline composition graph

solution for data pipeline composition. The algorithm 2 is comprised of two basic steps. First, it provides graph planner an initial graph I_S with a set of tasks and an initial set of constraints. Afterwards, it expands the solution graph based on the quality model constraints for each sub-goal which may include a composition solution. Secondly, if there is a solution available, the algorithm extracts it from the graph.

The composer initiates the planning process with line 1 that provides graph planner four inputs: (1) index i for the graph layer. (2) MT a master constraint table which includes the set of actions and respective constraints for each sub-goals. (3) an initial set of constraint C_{s0} defined by the composer. (4) an initial graph G layer for the corresponding tasks. In lines 2–7, with the initial input, the algorithm starts to expand the graph for the current layer, i.e., sub-goal. For each sub-goal, the planner extracts the actions and corresponding constraints. The graph continues to expand, if all of the constraints are satisfied; otherwise, there is no solution and the process is terminated. Meanwhile, after a solution is found for the current layer, the algorithm repeats the process for next layer based on the respective actions and constraints from master table. The graph is further expanded by adding a solution for each layer. Finally, lines 8 to 10 set the termination condition by verifying the solution graph size after each layer processing; if the size of the graph does not change, the algorithm returns failure and aborts the whole composition process. Lastly, line 11 returns the composition graph plan.

Service composition plan generation

The motivating scenario is used for Facebook service data pipeline composition as an illustration (see Fig. 3). The graph planner starts the planning process by initiating the graph with four inputs: graph layer index i, constraint table MT, initial constraint C_{s0} and graph G. C_{s0} defines the four initial constraints abstracted from the end user input: number of social information services required for analysis, temporal and spatial parameters (if any), minimum data volume and richness for the given topic, required sentiment analysis. MT contains four sets of constraints including pre and post conditions for four sub-goals based on service quality model: Data D, Preprocessing P, Location L, Analysis A. For instance, for the composition of a data pipeline for Facebook service, the sub-goal of data collection creates action ($FBServiceSelection \rightarrow G_{DCS}$) a data gathering service is allocated. To successfully complete the current state transition, two data constraints: data volume ($D : V_N$) and data richness ($D : R_{Sen}$) are validated, before proceeding

Table 1 Summary of dataset

Social information service	Collected data
Twitter (Tweets)	133043
Facebook (Comments)	108756
Youtube (Comments)	3983
Total Data Items:	245782

to next sub-goal. After the successful validation of D constraints, the data assessment action ($DataAssessment \rightarrow G_{DCS}$) is created to extract and assess the data features.

Secondly, for data preprocessing G_{PRS} sub-goal, preprocessing P constraints: data freshness ($P : F_R$) and data relevance ($P : D_{Rel}$) are validated. For ($P : F_R$), the irrelevant temporal and extra data filters are selected by creating the action ($FilterSelection \rightarrow G_{PRS}$). For ($P : D_{Rel}$), the subsequent action is initiated as ($IRServiceSelection \rightarrow G_{PRS}$) to choose an information retrieval service and state is transitioned. In addition, for sub-goal G_{LES}, a simple condition ($L : G_{DCS}$) is validated; whether a social information service is marked *True* or *False* for their geo-tagging facility. For instance, as Facebook does not provide geo-tagged data, thus location extraction task is created for ($LocationServiceSelection \rightarrow G_{LES}$) and state transitioned into next sate G_{DAS}. Finally, the analysis constraint: text type ($A : T_{Type}$) is used to create data analysis service selection task as ($AnalysisServiceSelection \rightarrow G_{LES}$). During the selection of a service at any sub-goal, if a required service is not selected, the planner aborts the composition process.

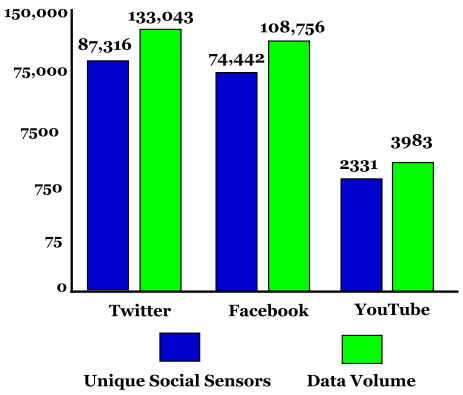
Experiments and evaluation

To evaluate the performance of BSDaaS, experiments are conducted on real-world dataset. A prototype is implemented and the performance of the proposed framework is evaluated by using the motivating scenario. A three fold sets of experiments are designed: (1) To evaluate the performance of data collection, preprocessing and location extraction components. (2) To evaluate the effectiveness of data analysis component by using three sentiment analysis services with human annotated data. (3) To provide the details of cost analysis for data pipeline integration. The prototype is developed in .Net Framework (2015) by using ASP.Net/C#. The experiments are conducted on a 3.40 GHZ Core i7 processor and 8 GB RAM on Windows 7.

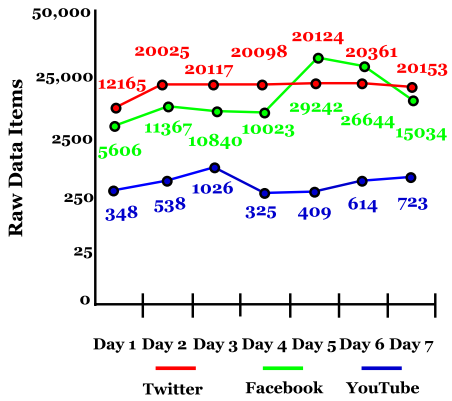
Dataset

The data is collected from three different social information services: Facebook, Twitter, and Youtube. All three services belong to different genre and provide various functions for the data sharing. The data is collected by using online services. For Facebook and Twitter, Facepager[8] is used for the data collection. For Twitter, Facepager uses 'Twitter Streaming API' to connect with Twitter stream. It gathers random tweets for the end user provided set of keyword(s). Moreover, the end user can also provide the geo-coordinates to gather tweets based on a specific geographical location. For Facebook, Facepager uses the Facebook provided 'Graph API' to collect Facebook posts and comments

[8] https://github.com/strohne/Facepager.



(a)



(b)

Fig. 4 **a** Richness: social sensors to data volume comparison **b** Freshness: data time-line

from public Facebook pages. However, unlike Twitter the 'Graph API' does not provide the facility to collect the data based on geographical locations. For Youtube, an online scraper service[9] is used to collect the comments from the videos.

For the data collection, the American president 'Donald Trump' is chosen as a topic of interest for online political discussion. The data is collected from 19-March-2017 to 26-March-2017. Random tweets from Twitter are collected by connecting to Twitter stream. For Facebook, the comments are collected from the official Facebook page[10] of 'Donald Trump'. Finally, for Youtube the comments are collected from the top ranked videos posted by social sensors. Table 1 presents the details of the data collected from three social information services.

Evaluation of system components

First, three properties: data volume, data richness, and data freshness, are used to evaluate the data collection component. Secondly, the performance of data preprocessor and location extraction component is evaluated.

[9] http://ytcomments.klostermann.ca/.

[10] https://www.facebook.com/DonaldTrump/.

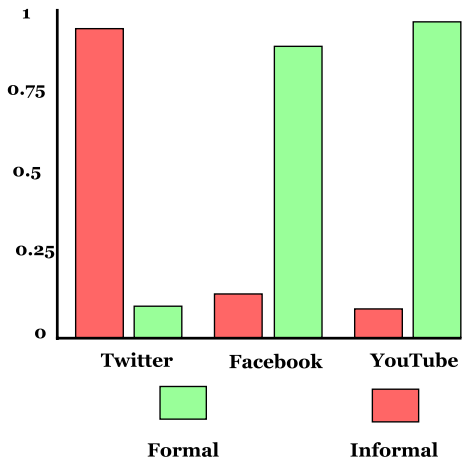


Fig. 5 Text classification results

The data volume V_N is measured by the data items collected from each social information service. For data richness R_{Sen}, unique social sensors are identified based on platform generated IDs. Figure 4a provides the ratio of unique social sensors with respect to data volume. For data freshness F_R, the collected data is sorted by the time and date. Initially, the data collection dates are used as temporal bounds. The data that does not comply with temporal requirements is discarded. Figure 4b presents the data time-line with 7 days of interval.

After the data collection, the data quality features are extracted from a random sample of collected dataset. First, the data relevance D_{Rel} is calculated. The data relevance for Facebook, Youtube and Twitter is 0.21, 0.25, 0.5 , respectively. In order to gain maximum information, boolean search model is implied based on 'search and stop terms' for three services. The search terms are used to include the data based on the relevant keywords, whereas, stop words or phrases are used to exclude the data that is irrelevant to the context. For instance, the tweet: "Angry Ivanka Trump Walks Out Of Cosmo Interview youtu.be/nKxxlftcYyY via @YouTube" contains the keyword 'Trump'. However, it is not relevant to the required context of 'President Trump'. Thus, 'Ivanka' can be used as a stop word to exclude such data items. For the filtering process, three keywords are utilized: 'US President', 'Trump', 'Donald Trump', as search terms. In contrast, three stop words: 'Ivanka', 'Obama', 'Hillary', for discarding the irrelevant data are applied. In current scenario, it is assumed that the search and stop terms are provided by the end users.

Secondly, the text type T_{Type} of the sample dataset is calculated. The classifier computes the results of *Formal to Informal Ratio* for Facebook, Twitter and Youtube. The classification results of formal and informal data classification of three social information services is presented in Fig. 5. Twitter has 8.9% of formal data and 91.1% of informal data. In comparison, Facebook and Youtube have 89.2% and 91.8% of formal data, respectively. Only 10.8% and 8.2% of data for Facebook and Youtube is classified as informal.

Table 2 Evaluation of data filtering results

Social information services	Language based filter	Search terms based filter	Stop words based filter	Location base filter
Twitter (SNR)	6.01	1.17	5.66	0.01
Facebook (SNR)	7.81	0.63	3.54	0.11
Youtube (SNR)	78.66	0.52	3.59	0.13

For the location extraction, Facebook and Youtube do not provide the geo-tagged data for public access, thus the geo-tagging condition is determined as *False*. To extract geo-locations from Facebook and Youtube, the Stanford NER[11] (Named Entity Recognizer) is used. Although the NER library parses the text and abstracts the location names (e.g., cities), yet it is a probabilistic approach to retrieve the social sensor geo-location. In current scenario, it is considered that the location appears with the conjunction of two words 'in' and/or 'from' as social sensor location. However, more sophisticated semantic methods are required to be investigated to detect actual locations of social sensors. In contrast, Twitter provides the geo-tagged data for public access. Hence, the geo-tagging condition for Twitter is computed as *True*. For Twitter, all the non-geo-tagged data is filtered out.

To evaluate the performance of the data preprocessing, *Signal-to-Noise Ratio (SNR)* is used as an evaluation metric. In each data filtering step, *Signal* is the number of relevant data items remaining, and the *Total* is the number of all items for each social information service. For language based filter, the number of initial data items are considered as *Total* in the dataset for each social information service.

$$SNR = \frac{Signal}{Noise} = \frac{Signal}{Total - Signal} \tag{9}$$

Table 2 presents the results[12] of SNR for each data filtering step. In this paper, all non-English data is filtered out. It can be observed from Table 2 that language based filter has the highest SNR. In search terms based filter the SNR significantly decreases. However, in stop words based filter the SNR again exceeds 1. Finally, the location extraction filter shows the lowest SNR among all of the filters.

Evaluation of data analysis component

For the evaluation of data analysis component, 'Black Box Testing' method is applied. For sentiment analysis, three sentiment analysis services are employed: Alchemy-API, Microsoft Text Analytics API, and Senti-Strength. Former two APIs are used to analyze formal text based on blogs, social information services, articles, etc. While later is designed to analyze the informal text (i.e., Twitter).

For evaluation, first, a sub-set of each social information service is manually annotated by human users into three categories: Positive, Negative, and Neutral. Based on the extracted text type from the sample datasets, Facebook and Youtube data is analyzed

[11] http://nlp.stanford.edu/software/CRF-NER.shtml.

[12] For the sake of simplicity, we have combined the extra data filter with stop words based filter.

Table 3 Evaluation of data analysis results

	Accuracy	Positive recall	Negative recall	Neutral recall	Positive precision	Negative precision	Neutral precision
Twitter	69.01%	85.71%	62.26%	90.91%	37.5%	100%	45.45%
Facebook	66.2%	74.28%	81.82%	28.57%	83.87%	50%	75%
Youtube	63.23%	86.36%	55.55%	40%	57.57%	66.67%	80%
–							
BSDaaS	66.19%	79.68%	63.96%	51.43%	63.75%	71.72%	54.84%
SVM	33.33%	68.75%	32%	5.26%	30.55%	47.01%	14.28%
NB	38.33%	25%	20%	73.68%	50%	41.67%	35%
Dictionary	46.67%	87.5%	48%	10.52%	41.18%	75%	20%

with Alchemy-API and Microsoft Text Analytics API, respectively. Twitter data is analyzed with Senti-Strength. Secondly, three evaluation metrics are considered: Accuracy *AC*, Precision *PP*, Recall *PR*. The evaluation metrics is defined below:

- Accuracy $AC = \frac{AR}{TR}$, TR is the number of all data items and AR shows correctly classified reviews, e.g., positive.
- Recall $PR = \frac{PC}{TP}$, calculates the accuracy of one type of reviews, e.g., positive. TP is the number of all positive reviews and PC is the number of correctly classified positive reviews.
- Precision $PP = \frac{PC}{PC+PW}$, calculates only one type of correct classification, e.g., positive. PC is the number of originally labeled positive reviews and PW shows the reviews that wrongly classified as positive.

Finally, to compare the efficiency of our approach, the annotated dataset is compared with three different sentiment analysis approaches: Support Vector Machine (SVM), Naive Bayes (NB), and Dictionary Based Approach. The former two approaches are pretrained with the baseline datasets,[13] whereas, the later is comprised of semantic corpus based on TextBlob [54]. *It is important to mention here that the three analysis services and comparison approaches used in evaluation are treated as black boxes.*

Table 3 provides the evaluation details of three individual sentiment analysis services including average results of BSDaaS and comparison approaches. With in the BSDaaD composition, the highest accuracy 69.01% is achieved from Senti-Strength. On the other hand, Microsoft Text Analytics API has the lowest accuracy of 63.23%. Since Senti-Strength is designed specifically for informal text, thus it has managed to achieve better results in terms of recall measures and negative precision. However, it shows less effective results in terms of positive and neutral precision. Both sentiment analysis services used for Facebook and Youtube have shown mixed results with various variations. For instance, both sentiment analysis services have a notable disparity in negative recall and positive precision. In addition, due to a smaller number of data items being labeled as neutral, the neutral precision is higher; whereas the neutral recall remains significantly lower.

[13] http://www.nltk.org/.

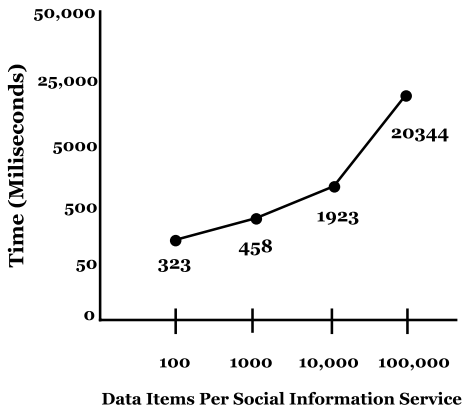


Fig. 6 Average data integration throughput

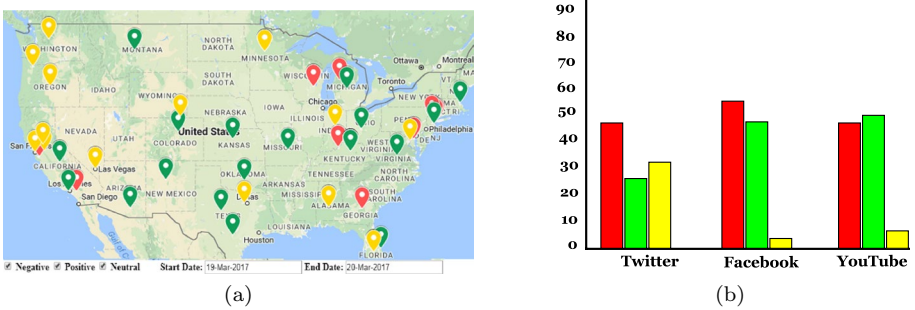


Fig. 7 **a** Sentiment detection on map **b** Composite sentiment analysis results

One reason for such variation in results is due to the length of the text of Facebook and Youtube comments. Social sensors tend to write longer comments with mixed opinion which include irony and sarcastic remarks. As a result, many data items are wrongly classified as negative or positive which leads to significant variation in the performance. In contrast, as the models of comparative approaches are not trained and validated with the current datasets, the performance remains insignificant and only exceeds in negative precision and positive recall. Eventually, the outcome of comparison points out the limitations of traditional sentiment analysis approaches and strengthens our hypothesis of services composition based on social information service quality features.

Evaluation of data integration component

The data analysis services used for each data pipeline, i.e., dataset, provide results in heterogeneous formats. For instance, Senti-Strength rates the polarity of negative and positive sentiment at the scale of (± 1 to ± 5). In contrast, Microsoft Text Analytics API grades the sentiment polarity from (0% to 100%). Where the score is less than 50% it is categorized negative, and when the score is greater than 50% it is classified as positive. Thus, prior to the final integration, the analysis results of each dataset is normalized. The

following function for the normalization is used which converts the data to a scale of 0 to 1.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{10}$$

where $x = (x1, ..., xn)$ are the polarity scores of sentiment and z_i is the ith normalized value. To evaluate the data normalization and final integration performance, *Throughput* is exploited as an evaluation metric. *Throughput* equals the total number of data items processed by the data integrator component. Figure 6 shows the average throughput time for the three social information service data integrations.

For the data visualization, a prototype of the BSDaaS system is implemented to provide the sentiment analysis results. 'USA' is defined as a spatial parameter, and data collection dates as temporal parameters. Figure 7a shows the sentiment by using a live Google map based on social sensor geo-locations. An end user can view the social sensor locations based on various type of sentiment between given dates. In Fig. 7b, the overall sentiment analysis results are composed in a percentage bar chart for three social information services. It is observed that overall social sensors have negative sentiment for US president. Social sensors on Twitter have highest percentage of negative sentiment, followed by second highest percentage of neutral sentiment. While, Twitter has lowest positive sentiment for US president. In comparison, social sensors on Facebook and Youtube have almost equal percentage of negative and positive sentiment. Social sensors on Facebook have slightly higher degree of negative sentiment. Meanwhile, social sensors on Youtube have marginally higher positive sentiment for US president. However, only a small percentage of social sensors have neutral sentiment on Facebook and Youtube.

Conclusions and future work

In this paper, a service composition framework is defined and implemented which extracts big social data from multiple social information services, and analyzes and transforms it into meaningful information based on their different data features. The framework established a novel service quality model that captures the heterogeneous features of social information services. The framework also includes a data pipeline infrastructure that applies the service quality features to compose multiple services for various social information services. In addition, the developed framework dynamically extracts and assesses the quality features of social information services and composes appropriate services required for various information analysis. In addition, a quality model driven service composition technique based on graph-planning is developed. A prototype of the framework is implemented and experiments are conducted on real dataset. Finally, the evaluation of results is demonstrated to show the efficiency of the framework in comparison with the existing techniques.

Future work As part of future work, we are interested to extend the data quality model. Currently, we intentionally ignored the demographic features (e.g., age, race) of social sensors as data quality features. In addition, we only used the data from three social

information services and only consider the data (i.e., text) written in English language. In contrast, the cyberspace of social information services is large and contains various natural languages. Incorporating the analysis of different natural languages and a larger pool of social information services into the proposed model is part of the future work.

Acknowledgements
Not applicable.

Author contributions
KA performed the primary literature review, data collection and analysis for the research work. Manuscript was drafted by KA, MH, CT and XZ. KA conceptualized the research idea to other authors. MA and CT coordinated the research process for completing the manuscript. All authors worked as a team to formulate the framework and methodology, and reviewed the article. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets collected and analysed during the current study are not publicly available due to process of the data collection which was performed privately/independently, and datasets were not shared with anyone but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 January 2021 Accepted: 2 May 2022
Published online: 15 May 2022

References

1. Becker D, King T, McMullen B. Big data, big data quality problem. In: IEEE International Conference on Big Data. 2015; pp. 2644–2653.
2. Ahsaan S, Mourya A. Big data analytics: challenges and technologies. Ann Faculty Eng Hunedoara. 2019;17(4):75–9.
3. Abdrabo M, Elmogy M, Eltaweel G, Barakat S. Enhancing big data value using knowledge discovery techniques. Inf Technol Comput Sci. 2016; 1–12.
4. Takeshi S, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: 19th International Conference on World Wide Web, ACM. 2010; pp. 851–860.
5. Kaplan A, Haenlein M. Users of the world, unite! the challenges and opportunities of social media. Bus Horiz. 2010;53:59–68.
6. Musaev A, Wang D, Pu C. Landslide detection service based on composition of physical and social information services. In: IEEE International Conference on Web Services.2014; pp. 97–104.
7. El Alaoui I, Gahi Y. The impact of big data quality on sentiment analysis approaches. Proc Comput Sci. 2019;160:803–10.
8. Nilashi M, Minaei Bidgoli B, Alrizq M, Alghamdi A, Alsulami A, Samad S, Mohd S. An analytical approach for big social data analysis for customer decision-making in eco-friendly hotels. Expert Syst Appl. 2021; 186.
9. Singh T, Kumari M. Burst: real-time events burst detection in social text stream. J Supercomput. 2021;77(10):11228–56.
10. Ali K, Hamilton M, Thevathayan C, Zhang X. Social information services: a service oriented analysis of social media. In: International Conference on Web Services. 2018; pp. 63–279.
11. Bebić D, Volarevic M. Do not mess with a meme: the use of viral content in communicating politics. Commun Soc. 2018;31(3):43–56.
12. Kumar R, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl-Based Syst. 2015;89:14–46.
13. Dai S, Gao Q, Fan Z, Kang G. User perceived quality of online social information services: from the perspective of knowledge management. In: IEEE International Conference on Industrial Engineering and Engineering Management. 2007; pp. 482–486.
14. Ali K, Dong H, Bouguettaya A, Hadjidj R. Sentiment analysis as a service: a social media based sentiment analysis framework. In: International Conference on Web Services.2017; pp. 660–667.
15. Wan S, Paris C. Improving government services with social media feedback. In: Proceedings of the 19th International Conference on Intelligent User Interfaces.2014; pp. 27–36.

16. Tinoco F, Hernández G, Zepahua J, Zepahua B, Mazahua L. A brief review on the use of sentiment analysis approaches in social networks. In: International Conference on Software Process Improvement. 2017; pp. 263–273.
17. Musaev A, Wang D, Calton P. Litmus: a multi-service composition system for landslide detection. IEEE Trans Serv Comput. 2015;8:715–26.
18. Thelwall M, Buckley K, Cai D, Kappas A. Sentiment strength detection in short informal text. J Am Soc Inform Sci Technol. 2010;61:2544–58.
19. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 2014;5:1093–113.
20. Cuomo M, Tortora D, Foroudi P, Giordano A, Festa G, Metallo G. Digital transformation and tourist experience co-design: big social data for planning cultural tourism. Technol Forecasting Soc Change. 2021; 162.
21. Cheung M, Pires G, Rosenberger III P, Leung W, Chang M. The role of social media elements in driving co-creation and engagement. Pacific J Mark Logist. 2021.
22. Fujiwara T, Müller K, Schwarz C. National bureau of economic research. Pacific J Mark Logist. 2021; 28849.
23. Zhou X, Chen L. Event detection over twitter social media streams. The VLDB J-Int J Very Large Data Bases. 2014;23(3):381–400.
24. Kitazawa K, Hale S. Social media and early warning systems for natural disasters: A case study of typhoon etau in Japan. Int J Disaster Risk Reduction. 2021; 51
25. Barbara M, Manso M. The role of social media in crisis. In: International Command and Control Research and Technology Symposium. 2012; pp. 19–21.
26. Finau G, Tarai J, Varea R, Titifanue J, Kant R, Cox J. Social media and disaster communication: a case study of cyclone Winston. Pac J Rev. 2018;24(1):123–37.
27. Boghiu S, Gîfu D. A spatial-temporal model for event detection in social media. Proc Comput Sci. 2020;176:541–50.
28. Phengsuwan J, Shah T, Thekkummal N, Wen Z, Sun R, Pullarkatt D, Ranjan R. Use of social media data in disaster management: a survey. Future Internet. 2021;13(2):46.
29. Wang Z, Ye X. Social media analytics for natural disaster management. Int J Geogr Inf Sci. 2018;32(1):49–72.
30. Kankanamge N, Yigitcanlar T, Goonetilleke A, Kamruzzaman M. Determining disaster severity through social media analysis: Testing the methodology with south east queensland flood tweets. Int J Disaster Risk Reduction. 2020; 42.
31. Patil H, Atique M. Sentiment analysis for social media: a survey. In: 2nd International Conference on Information Science and Security (ICISS).2015; pp. 1–4.
32. Serrano-Guerrero J, Olivas J, Romero F, Herrera-Viedma E. Sentiment analysis: a review and comparative analysis of web services. Inf Sci. 2015;311:18–38.
33. Keith Norambuena B, Lettura E, Villegas C. Sentiment analysis and opinion mining applied to scientific paper reviews. Intel Data Anal. 2019;23(1):191–214.
34. Guellil I, Boukhalfa K. Social big data mining: a survey focused on opinion mining and sentiments analysis. In: 12th International Symposium on Programming and Systems (ISPS). 2015; pp. 1–10.
35. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl-Based Syst. 2021; 226.
36. Mehta P, Pandya S. A review on sentiment analysis methodologies, practices and applications. Int J Sci Technol Res. 2020;9(2):601–9.
37. Niknejad N, Ismail W, Ghani I, Nazari B, Bahari M. Understanding service-oriented architecture (soa): a systematic literature review and directions for further investigation. Inf Syst. 2020; 91.
38. Hammoudeh M, Epiphaniou G, Belguith S, Unal D, Adebisi B, Baker T, Watters P. A service-oriented approach for sensing in the internet of things: intelligent transportation systems and privacy use cases. IEEE Sens J. 2020;21(14):15753–61.
39. Hustad E, Olsen D. Creating a sustainable digital infrastructure: the role of service-oriented architecture. Proc Comput Sci. 2021;181:597–604.
40. Hayyolalam V, Kazem A. A systematic literature review on qos-aware service composition and selection in cloud environment. J Netw Comput Appl. 2018;110:52–74.
41. Wang J, Yang Y, Wang T, Sherratt R, Zhang J. Big data service architecture: a survey. J Internet Technol. 2020;21(2):393–405.
42. Saggi M, Jain S. A survey towards an integration of big data analytics to big insights for value-creation. Inf Process Manage. 2018;54(5):758–90.
43. Neves P, Schmerl B, Cámara J, Bernardino J. Big data in cloud computing: features and issues. IoTBD. 2016; 307–314.
44. Ming Z, Kumar A, Ali M, Chong P. A cloud-based network architecture for big data services. In: 14th International Conference on Pervasive Intelligence and Computing.2016; pp. 654–659.
45. Vu H, Asal R. A framework for big data as a service. In: IEEE International Conference on Digital Signal Processing (DSP). 2015; pp. 492–496.
46. Khan S, Shakil K, Ali S, Alam M. On designing a generic framework for big data-as-a-service. In: 1st International Conference on Advanced Research in Engineering Sciences (ARES), IEEE. 2018; pp. 1–5.
47. Persico V, Pescapé A, Picariello A, Sperlí G. Benchmarking big data architectures for social networks data processing using public cloud platforms. Futur Gener Comput Syst. 2018;89:98–109.
48. El Alaoui I, Gahi Y, Messoussi R, Chaabi Y, Todoskoff A, Kobi A. A novel adaptable approach for sentiment analysis on big social data. J Big Data. 2018;5(1):1–18.
49. Grossman A, Frieder O. Information retrieval: algorithms and heuristics. Sci Business Media. 2012; 15.
50. Kevin C, Potdar V, Dillon T. Content quality assessment related frameworks for social media. In: Computational Science and Its Applications-ICCSA.2009; pp. 791–805.
51. Aggarwal C, Abdelzaher T. Integrating sensors and social networks. Soc Netw Data Anal. 2011; 379–412.
52. Potthast M, Stein B, Loose F, Becker S. Information retrieval in the commentsphere. ACM Trans Intel Syst Technol (TIST). 2012;3(4):68.
53. Siriweera S, Paik I, Kumara B. Constraint-driven dynamic workflow for automation of big data analytics based on graphplan. In: International Conference on Web Services.2017; pp. 357–364.

54. Loria S, Keen P, Honnibal M, Yankovsky R, Karesh D, Dempsey E. TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/.

Publisher's Note