Sentiment Analysis as a Service

A thesis submitted for the degree of Doctor of Philosophy

Kashif Ali

Master of Science - Computer Software Engineering - National University of Sciences and Technology, Islamabad, Pakistan
Bachelor of Science - Computer Science - Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi, Pakistan

School of Science
College of Science, Engineering, and Health

RMIT University
Melbourne, Victoria, Australia

March 2019
Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Kashif Ali
School of Science
RMIT University
28 March 2019
Acknowledgments

This journey of my PhD has been an incredible experience for me, not only in developing the research skills, but also helped me to sharpen my personality.

First, I would like to sincerely thank my supervisory team, Associate Professor Margaret Hamilton, Senior Lecturer Dr. Charles Theyvatham and Associate Professor Xiuzhen (Jenny) Zhang, for their unlimited support and guidance during the pursuit of my PhD candidature.

Secondly, I am truly thankful to the School of Science at RMIT University, Australia, for their financial assistance through scholarship and travel grants which helped a lot to just focus on my research and academic activities without any financial worries.

I am also grateful to the staff members of College of Science, Engineering, and Health, and in particular to the School of Science, who assisted me in every possible way to resolve any administration issues throughout my candidature.

Finally, I am immensely thankful to my mother Ms. Wajida Shaheen, for her unlimited love, support and continuous prayers which never let me give up.
Dedication

I dedicate this thesis to my beloved mother Ms. Wajida Shaheen, a brave woman, who regardless of living in a developing country and facing countless economic and social challenges in her life, put me on the right path and motivated me to pursue my dreams through the power of knowledge and education.
List of Publications

Portions of the work presented in this thesis have previously appeared or are to appear in the following publications:


## Contents

Abstract .............................................................................................................. 2

1 Introduction .................................................................................................... 3

1.1 Motivation .................................................................................................. 4
1.2 The Problem Statement ............................................................................ 7
1.3 Limitations of Existing Solutions .............................................................. 9
1.4 Research Questions ................................................................................... 10
1.5 Contributions ............................................................................................. 12
1.6 Thesis Structure ......................................................................................... 14

2 Background .................................................................................................. 16

2.1 Social Information Services ..................................................................... 16
2.2 Social Information Service Applications ............................................... 18
2.3 Sentiment Analysis ................................................................................... 20

2.3.1 Sentiment Analysis Techniques ....................................................... 21
2.3.2 Spatio-Temporal Sentiment Analysis ............................................. 23

2.4 Service Oriented Architecture (SOA) ..................................................... 24

2.4.1 Cloud Computing ............................................................................... 26
2.4.2 Cloud Services and Models ............................................................... 27
2.4.3 Service Composition ......................................................................... 29

2.5 Big Data .................................................................................................... 31

2.5.1 Big Social Data ................................................................................ 32
2.5.2 Big Data and Cloud Services ............................................................ 33

2.6 Chapter Summary ...................................................................................... 33
## 3 Service Composition For Spatio-Temporal Sentiment Analysis

3.1 Introduction .................................................. 35
3.2 Motivating Scenario .......................................... 37
3.3 Related Work .................................................. 38
3.4 The Framework Overview .................................... 41
3.5 System Models ................................................ 42
  3.5.1 Component and Composite Service Model ............. 42
  3.5.2 Social Information Service Classification Model ..... 44
     Social Information Service Types ......................... 45
     Social Information Service Features .................... 46
  3.5.3 Social Information Service Quality Model ............ 47
3.6 Dynamic Service Composition Approach .................... 49
  3.6.1 Human Centric Tag Based Semantic Model .......... 49
  3.6.2 Social Information Service Classification Based Concept Tagging .... 50
  3.6.3 Classification-Driven Service Composition .......... 51
3.7 Experiments and Evaluation of Proposed Framework ......... 53
  3.7.1 Experimental Setup .................................... 54
  3.7.2 Health Surveillance: Case Study .................... 54
     Data Collection ........................................... 55
     Noise Removal .......................................... 55
     Geo-Tagging ............................................ 56
     Sentiment Analysis ..................................... 57
     Prototype Results ...................................... 57
  3.7.3 Evaluation of Service Composition Approach ........ 59
3.8 Chapter Summary ............................................. 60

## 4 Social Information Services: The Service-Orientation of Social Media

4.1 Introduction .................................................. 62
4.2 Motivating Scenario .......................................... 63
4.3 Related Work ................................................ 65
4.4 Methodology .................................................. 68
4.5 Abstraction: Social Information Services ................. 70
  4.5.1 Meta-Service Model .................................. 70
  4.5.2 Functional Model ...................................... 71
CONTENTS

4.5.3 Quality of Service Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
4.6 Conceptualization: Functional and Non-Functional Analysis . . . . . . . . 73
  4.6.1 Functional Properties Analysis . . . . . . . . . . . . . . . . . . . . . . . 73
  4.6.2 Non-Functional Properties Analysis . . . . . . . . . . . . . . . . . . . 74
4.7 Implementation: Social Information Service . . . . . . . . . . . . . . . . . . 80
  4.7.1 Service Profile . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
  4.7.2 Service Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
  4.7.3 Service Grounding . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
4.8 Chapter Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

5 QoS Driven Service Composition For Big Social Data 88
  5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
  5.2 Motivating Scenario . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
  5.3 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
  5.4 Extended Sentiment Analysis as a Service Framework . . . . . . . . . . . . 95
    5.4.1 Data Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
    5.4.2 Service Composition Layer . . . . . . . . . . . . . . . . . . . . . . . . 99
      QoS Model: Social Information Service . . . . . . . . . . . . . . . . . . . 100
      Quality of Service Driven Composition Approach . . . . . . . . . . . . . . 101
      Quality of Service Driven Composition Approach . . . . . . . . . . . . . . 103
  5.5 Experiments and Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . 107
    5.5.1 Evaluation of Data Collection Service Layer . . . . . . . . . . . . . . . 108
    5.5.2 Evaluation of Preprocessing and Location Extraction Service Layer . . 109
    5.5.3 Evaluation of Data Analysis Service Layer . . . . . . . . . . . . . . . . 113
    5.5.4 Evaluation of Data Integration Service Layer . . . . . . . . . . . . . . . 116
  5.6 Chapter Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118

6 Efficient Composition of Reusable Sentiment Analysis Results 120
  6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121
  6.2 Motivating Scenario . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 123
  6.3 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124
  6.4 The Solution Overview: Meta-Information Composition . . . . . . . . . . . 126
    6.4.1 Meta-Information Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
      Analysis Request Data Model . . . . . . . . . . . . . . . . . . . . . . . . . . 126
Analysis Result Data Model ........................................... 127

6.4.2 Service Request Validator ...................................... 129
6.4.3 Meta-Information Service Composition ..................... 130

6.5 Experiments and Evaluation ...................................... 131
6.5.1 Meta-Information Model Implementation ..................... 132
6.5.2 Meta-Information Composition Performance Evaluation .... 134

6.6 Conclusion .......................................................... 136

7 Conclusion ............................................................ 137
7.1 Summary of Contributions ........................................ 137
7.2 Future Research Directions ....................................... 140

Bibliography .......................................................... 143
List of Figures

1.1 Data Analysis Process ........................................... 4
1.2 Overview of Sentiment Analysis as a Service .................. 6

2.1 Service Oriented Architecture Components [Endrei et al., 2004] .................. 25
2.2 Cloud Service Layers ........................................... 28

3.1 Service Composition Framework For Sentiment Analysis ............ 38
3.2 Flow of Data Between System Components ........................ 41
3.3 Sentiment Analysis Service Composition Process .................. 52
3.4 Hayfever Surveillance Map: Detected Social Sensors With Hayfever .... 58
3.5 Aggregated Sentiment Analysis Results of Social Sensors .......... 58
3.6 Performance of Service Computation Algorithm .................. 59

4.1 Workflow Composition for Social Information Services Analysis ........ 64
4.2 The Service-Orientation Process of Social Information Services ........ 68
4.3 Irrelevant Data In Three Social Information Services by Topic ........ 75
4.4 Unique Social Sensors in Social Information Services by Topic ........ 76
4.5 Data Time-line in Social Information Services by Topic ............ 77
4.6 The OWL-S Ontology Structure ................................... 81
4.7 Service Profile Class ............................................ 82
4.8 Composite Service Process Model: Search Tweets .................. 83
4.9 Deployment Model For Service Groundings of Twitter ............... 84
4.10 Twitter Service Grounding For Twitter Streaming API ............... 85

5.1 Sentiment Analysis System Design Scenario ..................... 91
5.2 Architecture of Extended Sentiment Analysis as a Service Framework .... 94
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>The Facebook Data Pipeline Composition Using Graph Planning</td>
<td>106</td>
</tr>
<tr>
<td>5.4</td>
<td>Data Volume to Unique Social Sensor Ratio</td>
<td>108</td>
</tr>
<tr>
<td>5.5</td>
<td>(a) Throughput: Data Collection Services (b) Throughput: Unique Social Sensor Sorting</td>
<td>109</td>
</tr>
<tr>
<td>5.6</td>
<td>Freshness: Data Time-line of Social Information Service Data</td>
<td>110</td>
</tr>
<tr>
<td>5.7</td>
<td>Data Classification Results of Formal and Informal Text Types</td>
<td>111</td>
</tr>
<tr>
<td>5.8</td>
<td>Execution Time of Data Filtering Services</td>
<td>112</td>
</tr>
<tr>
<td>5.9</td>
<td>Throughput: Comparison of Data Analysis Services</td>
<td>113</td>
</tr>
<tr>
<td>5.10</td>
<td>Average Data Integration Throughput of Integrator</td>
<td>117</td>
</tr>
<tr>
<td>5.11</td>
<td>(a) Sentiment Visualization on Map (b) Composite Sentiment Analysis Results</td>
<td>118</td>
</tr>
<tr>
<td>6.1</td>
<td>Meta-Information Service Composition Scenario</td>
<td>123</td>
</tr>
<tr>
<td>6.2</td>
<td>Meta-Information Service Composition Scenario</td>
<td>128</td>
</tr>
<tr>
<td>6.3</td>
<td>Analysis Request Structure</td>
<td>133</td>
</tr>
<tr>
<td>6.4</td>
<td>Analysis Response Structure</td>
<td>133</td>
</tr>
<tr>
<td>6.5</td>
<td>Meta-Information Transformation Time</td>
<td>134</td>
</tr>
<tr>
<td>6.6</td>
<td>Meta-Information Composition Throughput Comparison</td>
<td>135</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Social Information Service Classification Model .......................... 46
3.2 Noise Filtering Results .................................................. 55

4.1 Dataset Details (By Topic of Interest) ................................. 70
4.2 Social Information Service Functional Classification ................. 71
4.3 Functional Analysis: Data Sharing of Social Information Services by Topic 74
4.4 Signal to Noise Ratio (SNR) of Spatial Information ...................... 78
4.5 Formal To Informal Data Ratio ........................................... 78
4.6 Lexical Analysis of Social Information Services by Topic ............... 79
4.7 Meta-Data Analysis of Social Information Services ..................... 80

5.1 Summary of QoS Model Notations ......................................... 100
5.2 Summary of Composition Notations ....................................... 103
5.3 Summary of Dataset: Collected Data Items ............................... 107
5.4 Evaluation of Data Filtering Results ..................................... 112
5.5 Evaluation of Data Analysis Results ..................................... 115

6.1 Summary of Meta-Information Models Notations ......................... 127
6.2 Summary of Dataset ....................................................... 132
Abstract

This research focuses on the design and development of a service composition based framework that enables the execution of services for social media based sentiment analysis. Our research develops novel analytical models, composition techniques and algorithms which use services as a mean for sentiment abstraction, processing and analysis from large scale social media data. Current sentiment analysis techniques require specialized skill of data science and machine learning. Moreover, traditional approaches rely on laborious and time-consuming activities such as manual dataset labelling, data model training and validation. This makes overall sentiment analysis process a challenging task. In comparison, services are ‘ready-made’ software solutions that can be composed on-demand for developing complex applications without indulging in the domain specific details. This thesis investigates a novel approach that transforms traditional social media based sentiment analysis process into a service composition driven solution.

In this thesis, we begin by developing a novel service framework that replaces the traditional sentiment analysis tasks with online services. Our framework includes a new service model to present services required for sentiment analysis. We develop a semantic service composition model and algorithm that dynamically composes various services for data collection, noise filtering and sentiment extraction. In particular, we focus on abstracting sentiment based on location and time. We then focus on enhancing the flexibility of our proposed service framework to compose appropriate sentiment analysis services for highly dynamic and changing features of social media platforms. In addition, we aim to efficiently process and analyze large scale social media data.

In order to enhance our service composition framework, we propose a novel approach to formalize social media platforms as cloud enabled services. We develop a functional and quality of service (QoS) model that captures various dynamic features of social media platforms. In addition, we devise a cloud based service model to access social media platforms as services
by using the Ontology Web Language for Service (OWL-S). Secondly, we integrate the QoS model into our existing composition framework. It enables our framework to dynamically assess the QoS of multiple social media platforms, and simultaneously compose appropriate services to extract, process, analyze and integrate the sentiment results from large scale data. Finally, we concentrate on efficient utilization of the sentiment analysis extracted from large scale data. We formulate a meta-information composition model that transforms and stores sentiment obtained from large streams of data into re-usable information. Later, the re-usable information is on-demand integrated and delivered to end users.

To demonstrate the performance and test the effectiveness of our proposed models, we develop prototypes to evaluate our composition framework. We design several scenarios and conduct a series of experiments using real-world social media datasets. We present the results and discuss the outcomes which validate the performance of our research.
Chapter 1

Introduction

Social media platforms (i.e., social information services) such as Facebook, Twitter, and YouTube have emerged as powerful and free public tools for online data sharing [Dingli et al., 2015; Baruah, 2012]. Social information service users utilize these platforms for various purposes such as personal, professional and leisure. Regardless of the usage intention of these services, the data shared by social information service users contains valuable information such as different opinions and sentiments [Pak and Paroubek, 2010a; Vinodhini and Chandrasekaran, 2012]. Social information service users continuously express their opinions on various topics of interests such as politics, health, and entertainment. In case of any sudden or unexpected event, these users act as information sensors, and start pushing large data streams in the forms of text, images, and videos, on social information services. Therefore, these users can be interpreted as social sensors [Takeshi et al., 2010]. In addition, social information services enable social sensors to embed their data with geographical locations by using geo-tagging facilities. As a result, geo-tagged data provides a promising opportunity to acquire insights of social sensors based on location and time (i.e., spatio-temporal) features [Hu et al., 2013a].

The increasing popularity of social information services has attracted researchers and practitioners to leverage their data for multiple applications [Maynard et al., 2012]. Specifically, there is a significant interest in harvesting opinions and sentiments from social information service data. However, the data is available in unstructured format. Transforming this data into useful information is a challenging task and requires various techniques. One mechanism to abstract meaningful information from textual data is sentiment analysis [Pang and Lee, 2008; Liu and Zhang, 2012]. A typical social information service based sentiment
analysis process comprises of three steps as shown in Figure 1.1. Firstly, the data is collected from various social information services. Secondly, the data is preprocessed by using various noise filtering steps and converted into analyzable format. Finally, the data is analyzed and the required information such as sentiment is extracted. There are several applications of sentiment analysis including market and FOREX rate prediction, box office prediction, business analytics, recommender systems, marketing intelligence and political crisis coverage [Ravi and Ravi, 2015].

In this chapter, we elaborate the motivation behind this research, with information about the issues and research challenges. We discuss the limitations of existing approaches. We define the research questions and how we will address them in the thesis. Finally, we highlight our contributions, and give an overview of the organization of the thesis chapters.

1.1 Motivation

Service oriented architecture (SOA) has emerged as a software design pattern where services are used as software components which are assembled to develop new applications [Brown et al., 2002]. In SOA, services are considered as building blocks for developing an application. A service is defined as a self-describing, discoverable and platform independent computational elements that can be accessed on network via its programmable interfaces [Papazoglou, 2003]. A SOA based application consists of one or more services. It is possible that these services are hosted in a distributed network at different physical locations. In order to formulate a single application in a distributed network, services communicate and exchange data through messages by using multiple protocols [Arsanjani, 2004]. Moreover, services are utilized as a delivery mechanism to transport the data to end users and legacy applications [Wu et al., 2007].

SOA provides several distinct features in comparison to traditional software development models. For example, SOA delivers a powerful abstraction that veils the implementation
details from the service end users and application developers [Neiat et al., 2014]. The end users are not concerned with how the data is stored and processed by a service. This enables end users to visualize a service as ‘black box’ with respect to inputs and outputs. Thus, end users focus on using services in different applications without knowing the underlying technical complexities. In addition, SOA enables to develop an application in loosely coupled layers [Papazoglou et al., 2007]. Each layer constitutes a set of component services capable of completing specific tasks. This layering mechanism allows developers to design their application with great flexibility. Another major benefit of SOA is the re-usability of services. Existing services are binded based on the application requirements [Pautasso and Alonso, 2005]. Similarly, as the requirements of end users change, new services can be rapidly added in the existing application to enhance its capabilities. Last but not least, services can be classified based on their functional and quality of service (QoS) or non-functional features (e.g., price, response time). Hence, an end user can develop an application by satisfying multiple functional and QoS requirements [Zhai et al., 2009]. As a result, the growth of SOA has facilitated in building complex applications without indulging into the particular domain specific technical details.

Cloud computing has originated as an evolutionary form of SOA. Cloud computing provides a computational model where resources such as software and hardware are delivered as services that can be supplied to end users through Internet in an on-demand fashion [Zhang et al., 2010]. Cloud computing leverages from the design principle characteristics of ‘Service-Orientation’ which makes a reciprocal relationship between service oriented computing and cloud computing [Wei and Blake, 2010]. With the flexibility of leasing resources as services, cloud computing has given birth to the ‘as a service’ model. The ‘as a service’ model allows various service providers to develop and deploy their services in a cloud environment [Buyya et al., 2009]. On the other hand, end users can discover and compose these services as an application based on their requirements.

In our research, we propose a service-oriented sentiment analysis framework that is established on the principles of service oriented paradigm. We aim to harness the power and the ability to dynamically compose cloud services for abstracting sentiment from social information services. In this regard, we formulate the sentiment analysis processes into a service based architecture. In this architecture, the traditional processes required for sentiment analysis are replaced with the cloud services. In addition, sentiment analysis results are delivered to end users as a service. Therefore, we term this service oriented approach for sentiment analysis as Sentiment Analysis as a Service - (SAaaS). The proposed framework is comprised
of multiple service layers as shown in Figure 1.2. Each service layer has a set of task-specific services provided by various providers. These services are composed based on end user requests. The details of the service layers is as follows:

- **Data Gathering Service Layer** offers a set of data gathering services (e.g., APIs, crawlers) which are required to collect data from multiple social information services based on the end user’s sentiment analysis query.

- **Data Preprocessing Service Layer** receives the raw data from the data gathering service layer and removes noise by using different types of preprocessing services.

- **Information Extraction Service Layer** obtains the refined data from previous layer and extract the required information such as sentiments and opinions. In particular, we focus on abstracting the information based on spatio-temporal features. Finally, it delivers the extracted results into different presentation formats alike maps, charts, and tables.
CHAPTER 1. INTRODUCTION

1.2 The Problem Statement

A SOA based application requires resources such as software components to be exposed as services. In order to access a service, it should be discoverable and must have an interface [Elfatatry, 2007]. As a result, an application can discover and compose a set of services based on functional and QoS requirements. Similarly, the SAaaS approach necessitates the resources required for sentiment analysis to be exposed, discovered and composed as services in a distributed environment. In this regard, we have identified following research challenges:

- **Efficient Multi-Layered Service Model For Sentiment Analysis**: Sentiment analysis is a multi-step problem that requires data gathering, data preprocessing and information analysis. Each step is further divided into various sub-steps. For instance, there are multiple data collection services available online to capture data from social information services [Lomborg and Bechmann, 2014]. Likewise, there are multiple techniques available for filtering relevant information from collected data [Grossman and Frieder, 2012]. These techniques are often applied in different combinations to extract relevant data. For more complex analysis scenarios, the utilization of various types of information extraction techniques are needed. Hence, an efficient and robust multi-layered service access model is required to formulate sentiment analysis steps as service layers, and sub-steps as component services. Moreover, there are different QoS attributes associated with component services, and these services are needed to be dynamically selected for sentiment analysis upon end user requests. Therefore, the QoS of each service layers must be explicitly captured.

- **Sentiment Analysis Service Composition Model**: The composition mechanism is a powerful means to aggregate a set of services. However, the service composition is a challenging task [Sivasubramanian et al., 2009]. Although the service composition is well established, the composition of sentiment analysis services pose many research challenges. In traditional service composition approaches, the user defined constraints and QoS play decisive role in service selection and composition [Tian et al., 2003; Rajendran et al., 2010]. In comparison, for sentiment analysis, services are also required to cater the diversity of social information services and process their highly dynamic data. Therefore, a new service composition model is required on the top of existing composition mechanisms to cope with the various types of social information services for sentiment analysis.
• Efficient Interpretation of Social Information Services Diversity: There are various types of social information services available online for data sharing. Their data sharing mechanisms and social sensor engagement vary by platforms [Dai et al., 2007]. For instance, Twitter restricts its users to write text messages with limited number of characters. Meanwhile, Facebook does not have such restriction. As a result, the data produced by these services has dissimilar features in terms of size, text length, and quality. Furthermore, these features may change with respect to domains or topics of interest such as sports, politics, etc. Current sentiment analysis approaches are mainly data-oriented that require extensive data training [Kontopoulos et al., 2013; Prabowo and Thelwall, 2009]. These approaches do not differentiate distinct features of various social information services. In order to efficiently analyze multiple social information services for sentiment analysis, it is essential to understand their diversity at service level. Thus, social information services are required to be formalized as cloud services with functional and QoS features. Also, it is necessary to analyze the impact of various domains (e.g., politics) on these features.

• On-Demand Service Composition For Large Scale Data Analysis: Social information services generate a large amount of data at an unprecedented speed [Tang and Liu, 2012]. The processing of increasing volumes of social information service data (i.e., big social data) is an extremely difficult task [Cambria et al., 2016]. The dynamic features such as level of noise, size, of different social information services make the analysis of big social data difficult. For instance, the data collected from Facebook and Twitter differs in terms of structure and quality [Ochieng et al., 2016]. Hence, the same noise filtering service may not be applicable to both platforms for relevant data extraction. Similarly, one information extraction or sentiment classification service may work better on one platform when compared to another. Adding more to the complexity, a service composition for a specific domain of entertainment, for example, may not be applicable to another domain such as finance. Thus, a flexible service composition framework is essential that assesses the diverse big social data features, and on-demand composes appropriate services for sentiment analysis based on dissimilar features.

• Efficient Re-usable Information Composition Model: The information extraction from big social data collected from social information services requires a dedicated amount of time and resources [Hashem et al., 2015]. With the arrival of distributed technologies (e.g., Hadoop), faster parallel processing of data has been achieved. Nevertheless, as
the volume of data increases, the cost of infrastructure resources in terms of storing and processing grows [Katal et al., 2013]. Public political sentiment analysis, a typical sentiment analysis scenario, requires continuous data collection and storage for a long duration before the information can be extracted and presented to the end users. Reducing the resource utilization and service composition cost (i.e., throughput) of such scenarios require efficient information composition model capable of transforming large amounts of data as re-usable information.

1.3 Limitations of Existing Solutions

In this section, we underline the limitations of the existing approaches for social media based sentiment analysis. These limitations are further elaborated in the respective related work sections of Chapters 3 to 6 concerning individual research questions in a chronological order.

- **Data Centric Sentiment Analysis**: There are mainly two types of sentiment analysis approaches: 1) Machine Learning Approach. 2) Lexicon Based Approach. The machine learning based techniques [Medhat et al., 2014] are generically established using statistical methods. In machine learning techniques, the data is collected from various sources and stored in a repository. Next, the collected data is divided into training and testing datasets. The data items in the datasets are then manually labelled with the relevant classes (e.g., positive, negative). Later, an algorithm driven classifier model is trained with the training dataset and its accuracy is validated with the testing dataset. This process is repeated until desired accuracy of classifier is achieved. On the other hand, lexicon based approaches [Annett and Kondrak, 2008] rely on developing special corpus or dictionaries by collecting and categorizing words and phrases which are used for semantic matching for sentiment classification. Despite the differences between these two methodologies, both approaches are time-consuming and comprise of various rigorous tasks (e.g., data collection, manual labeling, algorithm training) and require specialized skills of machine learning and data science [Li et al., 2010].

- **Commercial Tools Based Sentiment Analysis**: In contrast to data centric approaches, there are several online tools and applications available for social information services based analytics and sentiment analysis. Although some of these tools provide the facility for analyzing sentiments, the majority of tools focus on social information
service monitoring and general-purpose information search and analysis. Among these tools some are specifically designed to analyze data of a single social information service like Twitter or Facebook and do not allow to analyze a set of social information services in an integrated environment. As a result, end users have to simultaneously use multiple analysis tools in an ad-hoc manner to analyze social information services of their choice [Wan and Paris, 2014]. Using multiple tools simultaneously is time-consuming and provides inconsistent perspectives of social sensor data.

- **Lack of Cloud Services Utility in Big Social Data Analysis:** The emergence of cloud service infrastructure has provided a viable platform to develop automated processes including big data collection, manipulation, and analysis [Demchenko et al., 2013]. However, existing efforts for big data analysis are mostly focused on infrastructure modeling [Horey et al., 2012] and high level of cloud service conceptualization [Vu and Asal, 2015; Ming et al., 2016]. The proposed approaches describe the outline of big data processing components and their relationships in the cyberspace of cloud services at process levels, rather investigate their formulation and composition at service levels. Moreover, the suggested big data processing components are highlighted as conceptual ideas rather concrete implementations. Another limitation is that the proposed frameworks do not distinguish between big data and big social data. Consequently, there is no abstraction provided to separate big social data from big data and process it by utilizing cloud services for information analysis scenarios such as sentiment analysis and opinion mining.

### 1.4 Research Questions

The main research goal of this thesis is to develop a service-oriented framework that can be used as a substitute for the traditional sentiment analysis approaches. In this proposed framework, instead of using data centric approaches, third party services provided by different service providers are composed to perform sentiment analysis. The proposed framework shall provide the means of abstraction to encapsulate the technical complexities of various sentiment analysis processes. At the same time, it utilizes the service oriented architecture of cloud services to collect, process and deliver sentiment analysis to end users.

In order to achieve our goal and overcome the above challenges, we have established following research questions.
1) How to design a service composition framework capable of composing services for sentiment analysis?

In this question, we first address the challenge of devising a service based architecture that formulates the traditional sentiment analysis process into a service composition driven sentiment analysis process. In this regard, we aim to achieve two objectives: 1) Design a formal service model that represents the structures of composite and component services with their relevant features necessary for sentiment analysis. 2) Develop a service composition mechanism that is capable of composing services for various types of social information services. In particular, this question focuses on composing services for spatio-temporal sentiment analysis of social information services.

2) How does the framework recognize the dynamic changes to social information services affected by different domains?

In the first question, we ascertain the baseline sentiment analysis process as a service composition problem. However, the fluid nature of social information services and different domains such as politics, entertainment and health, constantly affect the data features (e.g., noise, text length). Consequently, compositions of services for sentiment analysis demand to cater the changing characteristics of the data. Thus, it is essential to devise techniques capable of interpreting and adapting to dynamic nature of social information services. In our second research question, we explore a novel approach to analyze and detect the variability of social information services. We develop a mechanism that formulates the dynamic changes of social information services in terms of functional and QoS models comprising of various properties. As a result, we aim to extend the existing framework that can cope with the dynamic changes in social information service features for composing sentiment analysis services.

3) How to efficiently analyze big social data for sentiment analysis with dynamic QoS features?

In the third question, we aim to develop a dynamic service composition framework that can handle on-demand requests by simultaneously composing services for processing and analyzing multiple social information services for sentiment analysis scenarios. In this context, we expand our initial service composition framework (designed in research question 1) with the integration of social information service QoS model (devised in research question 2). In this question, we first focus on how to dynamically assess the QoS features of different
social information services for a given sentiment analysis scenario. Secondly, we investigate how the evaluated QoS measures can be utilized to concurrently compose appropriate data collection, preprocessing and data analysis services for multiple social information services. Finally, this question also looks into the aggregation of sentiment analysis results produced by simultaneous service compositions.

4) How to effectively reuse the sentiment analysis results of big social data for reoccurring analysis scenarios?

Forming a service composition plan for sentiment analysis is a complex task. Secondly, the big social data processing via services is both resource and time consuming. The previous research questions establish a framework that provides multi social information service based sentiment analysis through service composition. In this research question, we extend our framework and explore an effective approach that is able to preserve the sentiment analysis results for reoccurring service composition scenarios. In this regard, the sentiment analysis results are modeled and stored as meta-information, and on-demand the results are composed for repeating queries. We aim to eliminate the need of re-initiating the service composition and resource utilization for big social data processing, in case of repeating sentiment analysis requests. Moreover, the proposed technique will enable to integrate the existing meta-information (i.e., sentiment analysis results) with new sentiment analysis outputs. The objective of this question is to develop an efficient technique for re-utilizing the previously extracted sentiment analysis results.

1.5 Contributions

In this section, we discuss the the details of our research contributions made in this thesis to address the research questions.

A Service Oriented Framework For Social Information Services Based Sentiment Analysis

In our first contribution, we devise a service-oriented framework that avoids the need for human involvement for performing social information services based sentiment analysis through service driven data collection, preprocessing, location extraction and data analysis. The main novelty of our framework is the formulation of a service oriented based solution which serves
CHAPTER 1. INTRODUCTION

as an alternative for the traditional data centric sentiment analysis methods. In our framework, we first develop a novel classification model for social information services based on their generic features. Secondly, we propose a semantic matching driven service composition mechanism to compose the appropriate sentiment analysis services based on the classification of social information services. Our framework allows users to collect and aggregate sentiments of social sensors from any combination of social information services without indulging in the complexities of typical sentiment analysis techniques. Moreover, the framework specifically focuses on the sentiment analysis to be performed based on social sensors time and location. The implementation of our framework for identifying the flu outbreak by using the spatio-temporal properties of social sensors sentiments shows the effectiveness of our approach.

Service Orientation Of Social Media

The second contribution is a service oriented modeling of social media. By using the service-orientation, we formalize social media platforms as conventional Web/cloud services (i.e., social information services). The novelty of utilizing service-orientation is that a social information service can be described by its functional and non-functional (i.e., QoS) properties. Moreover, these properties will also allow to reflect the dynamic characteristics (e.g., spatial, temporal) a social information service may offer. In addition, treating social media as a homogeneous data source rather heterogeneous services with varying features in the past prevented the dynamic selection, composition and analysis of social information services. We apply Ontology Web Language for Service (OWL-S) to describe a Web based access model for social information services. We test the proposed model by analyzing the functional and QoS features of three different types of social information services with respect to three different topics of interests. The results show that the service-oriented model is a feasible approach to capture the dynamic features of social information services.

A QoS-aware Service Composition Framework For Big Social Data Analysis

Our third contribution is a dynamic service selection and composition approach for extracting sentiment analysis from big social data collected from different social information services. The extraction of sentiments from exponentially growing big social data volume and dynamically changing features is a challenging task. In this contribution, we develop a novel approach which 'on-the-fly' assesses QoS features of big social data obtained from multiple
types of social information services, and simultaneously composes appropriate services re-
quired for sentiment analysis based on the extracted QoS features. The proposed approach
contains a new QoS-aware service composition algorithm which uses the Graph-Planning
technique. The Graph-Planning based composition algorithm provides the ability to validate
the selection of processing and analysis services based on QoS features of large streams of
social information services data. The experiments using the new technique on real datasets
and comparing it with existing sentiment analysis techniques demonstrates the efficacy of
the proposed approach.

An Efficient Information Model For Re-utilizing Sentiment Analysis Results

The fourth contribution is an information modeling mechanism that enables the reuse of
sentiment analysis results obtained from big social data. The novelty of this approach is the
utilization of meta-information modeling for transforming and storing the extracted senti-
ments as reusable information, which can be on-demand integrated and delivered as a service.
The proposed approach is suitable for sentiment analysis scenarios (e.g., customer sentiment
patterns detection) which require to store and analyze large amounts of data, and where the
similar sentiment analysis requests are frequently repeated. Such analysis scenarios demand
time and resource intensive computation for information extraction. The significance of our
proposed model is that it allows information to be reused or combined with newly processed
information to meet the required analysis needs at the same time reduces the computa-
tional costs. The experiments on real-world data demonstrate the efficiency of our proposed
approach especially where sentiment analysis requests are heavily repeatedly.

1.6 Thesis Structure

The remainder of the thesis is organized as below:

In Chapter 2, we present a comprehensive review of core concepts included in this thesis.
We include social information services and the significance of its applications. We discuss the
area of sentiment analysis and its current approaches, and explain their particular character-
istics. In addition, we review the combination of spatio-temporal approaches used in social
information services based applications. We then review the domains of service composition
and cloud computing. Finally, we discuss the existing approaches for utilizing cloud services
for big social data analysis.

In Chapter 3, we present a service oriented framework that composes a series of services
for data collection, preprocessing and extracting sentiments based on spatio-temporal properties. It includes formal service models of component and composite services. It defines a classification model of social information services. We devise a classification model driven service composition mechanism for sentiment analysis. A prototype of the proposed approach is evaluated with a real dataset to show the applicability of our proposed framework.

In Chapter 4, we present a novel technique to formulate and analyze social media. We formally model social information service as cloud enabled services which are discoverable and accessible in Web environment by end users. We also propose functional and QoS models with respect to various features. The service orientation of social media provides the basis to manipulate, select and compose them in different analysis scenarios. We also present the experiment results and quantify the impacts of different domains on the proposed functional and QoS properties.

In Chapter 5, we extend our proposed service oriented framework with QoS driven service composition approach. In our initially proposed framework presented in Chapter 3, we used the social information service classification model to compose services for sentiment analysis. In this approach, we utilize the QoS model proposed in Chapter 4 to dynamically compose services for sentiment analysis of large-scale data. This extension of framework allows to simultaneously devise composition for a multi-platform analysis. We present the evaluation of the proposed approach through experiments.

In Chapter 6, we introduce a reusable information composition technique as an increment to our QoS driven service composition framework developed in Chapter 5. The proposed technique converts and stores the sentiment analysis of large streams of social information service as reusable information named as meta-information. The meta-information is composed on-demand and delivered as a service for reoccurring sentiment analysis queries. We devise a novel QoS model for meta-information composition service. Finally, we showcase the experiments and their results.

In Chapter 7, we present the summary of our all contributions and key findings. In addition, we discuss the potential future directions and some of the limitations of developed methods.
Chapter 2

Background

Our research work is focused at the intersection of the research areas of social information services, sentiment analysis and cloud service composition. To the best of our knowledge, there is no service composition framework for sentiment analysis in the literature. Therefore, this chapter provides the relevant constructs and the ideas of different research domains we leveraged in our work. In this chapter, we provide the background information and overview of concepts that form the foundations of our proposed framework. Section 2.1 covers the introduction and the earlier classifications of social information services. In Section 2.2, we review the application usage of social information services for different domains. We present the overview of sentiment analysis and its various techniques in Section 2.3. In addition, we provide a summary of spatio-temporal application with respect to social information service based sentiment analysis. Section 2.4 discusses the key concepts of SOA, cloud computing paradigms and models, and existing service composition techniques. The fundamentals of big data and previous studies of big data in the domain of cloud service computing is discussed in Section 2.5. Finally, the summary of this chapter is presented in Section 2.6.

2.1 Social Information Services

The term social information service is coined to define social media platforms (e.g., Twitter, Facebook, YouTube) as online services which provide information for different applications [Musaev et al., 2014]. On the other hand, the social media users are described as social sensors - who use social information services for information sharing. However, social media itself is not a new idea. The concept of social media originates from social networks where social groups are formed by members who have similar values and beliefs [Edosomwan et al.,
In recent times, the adaptation of digital media technologies (e.g., Web, Internet) has changed the dynamics of social networks into social media Web sites. The Oxford Dictionary defines social media as: “Web sites and applications that enable users to create and share content or to participate in social networking”. The Internet driven social media Web sites such as SixDegrees, AsianAvenue, and BlackPlanet started appearing in the 1990s [Boyd and Ellison, 2007]. In this research, we associate the terminology of Social Information Services to the notion of Social Media.

With the cheap availability of Internet access across the globe, social information services have gained a tremendous amount of popularity. Many social information services have surfaced on the Internet over time. These services encompass a wide range of online forums, blogs, chat rooms, file sharing platforms (including audio, video, images), networking, bookmarking sites [Mangold and Faulds, 2009; Hanna et al., 2011]. There is plethora of online accessible social information services by social sensors. However, these services have many distinct as well as overlapping functionalities. Social sensors choose social information services based on their functional features and personal interests [Kietzmann et al., 2011]. Based on the several similar features, the following six types of social information services [Kaplan and Haenlein, 2010] can be classified:

1. **Collaborative Projects:** enable the creation of content by social sensors in a combined and simultaneous effort. The groups of social sensors can contribute by adding, removing or updating content on collaborative Web-sites (e.g., Wikipedia).

2. **Blogs:** are the earliest and simplest form of social information services. Blogs such as TechCrunch are Web pages which are generally managed by one social sensor usually the owner. However, other social sensors can still interact through posting online comments at the end of the content.

3. **Content Communities:** are services (e.g., YouTube, Instagram) which allow social sensors to share contents (e.g., files, images, videos) with others.

4. **Social Networking Sites:** provide the ability to social sensors for creating personal profiles and inviting friends and family to access their information (e.g., text messages, videos, images). Social information services like Facebook and Myspace are prime examples of social networking sites.

---

1 https://en.oxforddictionaries.com/definition/social_media
5. **Virtual Game Worlds**: are online gaming services (e.g., World of Warcraft) in which social sensors appear as personalized avatars in a three dimensional environment and interact with others according to the rules defined for the game.

6. **Virtual Social Worlds**: are virtual worlds (e.g., Second Life) where social sensors live as “residents” by using digital avatars. They interact with or without any rules in a virtual three dimensional environment which simulates real life settings.

### 2.2 Social Information Service Applications

The increasing popularity and public nature of various types of social information services has successfully attracted a large number of social sensors. As a result, social information services provide an exponentially growing source of public data. This phenomenon has attracted people from academia and industry to extract information from the public data and utilize it in a range of application scenarios [Naaman, 2012]. Some applications of social information services in different domains are discussed below:

#### Business

Social information services are providing new avenues for business firms to satisfy their existing customers and help them acquire new ones. For instance, social information services help marketing and sales departments by providing a medium to promote products and services [Culnan et al., 2010]. Social sensors may share novel ideas and different opinions about products they consume which are valuable input for product manufacturers [Liu et al., 2013b]. For instance, Xiaomi, a mobile phone manufacturing company, modifies their new versions of cell phones based on social sensor discussion related to their products. Business competitors in the food industry are gaining a commercial edge by analyzing their communications with social sensors gathered from multiple social information services [He et al., 2013].

#### Civic

Social information services have become a prominent tool for the public voice. Social sensors point out and discuss social and public issues [Papadopoulos et al., 2012]. For instance, social sensors contemplate their countries’ political situation on different social information services which forecast the electoral results [Ceron et al., 2014]. Similarly, social sensors discuss the current ongoing policies and create collective actions such as signing petitions on
social information services [Obar et al., 2012]. Moreover, several government departments and agencies are taking initiatives to draw awareness and engage with public through different social information service based campaigns [Lee and Kwak, 2012].

Public Security

The easily accessible nature of social information services provide a possible source to track down criminals and violent people. For instance, social sensor issue alerts and warnings to others regarding an offense (e.g., car robberies) taking place [Featherstone, 2013]. At the same time, propagating these alerts can help to track offenders. In addition, the content analysis of social information services can help identify the radical behaviors and intentions of potential violent actions [Cohen et al., 2014]. Many governments and law enforcement agencies are now taking advantage of social information services to provide safety for their public by identifying malicious activities.

Crisis Management

Social information services are becoming an important instrument of communication for social sensors in emergency situations [Alexander, 2014]. During any natural disaster (e.g., earthquakes, floods), social sensors continuously share information by uploading text messages, videos, etc. This information helps authorities to locate missing people, and alert about the different aspects of the emergency [Hjorth and Kim, 2011]. In addition, social information services also lay the backbone for the crowd sourced based crisis mapping platforms built for disaster management [Gao et al., 2011]. The data produced by social sensors is collected from different social information services to carry out relief activities.

Media and Journalism

The age of the Internet and social information services has altered many traditional aspects of human life. One such impact can be observed on journalism. Social information services are helping to shape conventional methods of journalism such as print media and television broadcasting into citizen journalism [Hermida, 2012]. Social sensors are broadcasting news and reporting incidents by using the streaming capabilities of social information services [Poell and Borra, 2012]. Moreover, the growing consumption of social information services helps mainstream news organizations to expand their network coverage, as it can assist in gathering news and disseminating information from conflict zones [Ali and Fahmy, 2013].
Despite the several benefits and application scenarios of social information services, harvesting information from social information services is a complicated task. Social information services often contain a lot of noise and irrelevant data. In addition, the diversity among social information services and their data makes the information analysis more complex task.

2.3 Sentiment Analysis

Sentiment analysis is one of the mechanisms for extracting meaningful information from social information services. The domain of sentiment analysis is the intersection of three research disciplines: Information Retrieval, Natural Language Processing (NLP), and Artificial Intelligence [Schouten and Frasincar, 2016]. Sentiment analysis was considered as a classification problem where a document or piece of text is classified into three categories: Positive, Negative, and Neutral [Pang et al., 2002]. However, lately the term sentiment analysis has become synonymous for ‘opinion mining’. In broader context, sentiment analysis or opinion mining is considered as a task of extracting and analyzing subjective information such as sentiments, opinions, attitudes, emotions and appraisals towards different topics, products, individuals, entities, and services [Serrano-Guerrero et al., 2015].

The growth of social information services have influenced many small and big companies, and other organizations (e.g., governments) to adopt sentiment analysis on a commercial basis [Castellanos et al., 2011]. Their main aim is to inquire about what the public says about their policies, products and people. Consequently, the sentiment analysis helps them to convert public opinion into actionable insights and improve their competitiveness. Social information services based sentiment analysis has been used in many applications such as market and FOREX rate prediction [Ozturk and Ciftci, 2014], box office prediction [Jain, 2013], business intelligence [Chen et al., 2012], recommender systems [Zhang et al., 2014] and political sentiment analysis [Wang et al., 2012].

Sentiment analysis can be carried out on different levels. Sentiment analysis is further divided into four types of sub-problems [Feldman, 2013]. These types are explained below:

1. **Document Level Sentiment Analysis** is the simplest form of sentiment analysis where a document is assumed to have an opinion. Thus, the sentiment polarity of the document is calculated as a whole.

2. **Sentence Level Sentiment Analysis** considers that a document may contain multiple opinion about an entity by using a set of sentences. Hence, the document is split into
atomic sentences and sentiment polarity is extracted. This type of analysis aims to
determine sentiment at a fine-grained level.

3. **Aspect Level Sentiment Analysis** focuses on extracting the sentiment about an entity
   based on its different aspects (i.e., attributes). For instance, the aspect level senti-
   ment of a mobile phone would require extracting the sentiments about components like
camera, battery, color, etc.

4. **Comparative Sentiment Analysis** targets to discover opinions from a document which
   provide comparison between two or more entities (e.g., comparison between two mobile
   phones). Furthermore, the preferred entities are extracted from each opinion.

### 2.3.1 Sentiment Analysis Techniques

There are various techniques available for sentiment analysis. These techniques can be di-
vided into two main types: Machine Learning Methods, Lexicon Based Methods. There are
also hybrid techniques formed amalgamating the two main approaches [Maynard and Funk,
2011]. Although hybrid techniques combine the features of both sentiment analysis methods,
lexicons based approaches forms the core of hybrid techniques. In addition, both main senti-
ment analysis methods have their own advantages and disadvantages. Generally, a sentiment
analysis technique requires performing additional tasks known as feature selection [Medhat
et al., 2014]. A feature selection process may comprise of different steps such as parts of
speech (POS), stemming, term frequency count. In the remainder of this subsection, we give
an overview of both sentiment analysis methods.

**Machine Learning Methods**

Machine learning methods are also known as statistical methods for sentiment classification.
These methods consist of different algorithms that are trained by an example data (i.e.,
training data) gathered to learn the underlying patterns. In the example data, each instance
is labeled or provide with a class (e.g., positive, negative), and later, the trained algorithm
is used to classify an unlabeled data [Pak and Paroubek, 2010b], making it a classification
problem. The machine learning methods are further categorized into two sub-categories:
Supervised Learning, Semi or Unsupervised Learning.

The supervised learning methods are dependent on the existence of training documents
which are pre-labeled or annotated with different types of classes. Within the supervised
learning, four types of classifiers are used for sentiment detection.

1. **Probabilistic classifiers** predict the probability of data item as input belonging to a specific class. These classifiers are generally trained with different classes. Each class is a mix of terms, and the probability of a data item’s class is calculated by sampling the terms of training data. Commonly used probabilistic classifiers are Naive Bayes Classifier (NB) [Liu et al., 2013a] and Bayesian Network (BN) [Ortigosa-Hernández et al., 2012].

2. **Linear classifiers** convert an input into a vector space and predict the classification of the vector based on given linear combination of the training features or characteristics. There are many variations of linear classifiers; two of the most famous are Support Vector Machines (SVM) [Mullen and Collier, 2004], and Neural Network (NN) [Santos and Gatti, 2014].

3. **Decision tree classifiers** provide a hierarchical decomposition of the training or example data based on a conditional value [Quinlan, 1986]. The condition is usually occurrence or absence of one or more words. The division process is conducted recursively until the leaf nodes contain a minimum number of the records which determine the classification.

4. **Rule-based classifiers** define a set of rules for data space. The rules comprise of conditions applied on the feature set. Each condition has a respective class label as an output. The conditions are generally based on word(s) presence [Ma and Liu, 1998]. Rule-based classification models [Gilbert and Liu, 2014; Kim and Hovy, 2006] are developed with the conjunction of large lexical resources.

The supervised learning methods mainly rely on large training datasets. As a result, gathering large datasets and labeling the data items is a time consuming task. In comparison, there are several unsupervised and semi-supervised learning techniques available [Hu et al., 2013b; He and Zhou, 2011]. The unsupervised learning derives the classification by using set of rules and heuristics obtained from existing or previously known language knowledge [Balazs and Velsquez, 2016]. In contrast, semi-supervised learning methods incorporate a sentiment lexicon as a labeled feature set into the model classifier before predicting the class of unlabeled data items. Semi-supervised learning also encapsulates the self-training approach. In this approach, a classifier is first trained with a small set of labeled data, and then it is used to classify unlabeled data. The newly labeled values with high level of confidence index are then added to the training model and classifier is re-trained [Zhu, 2006].
CHAPTER 2. BACKGROUND

Lexicon Based Methods

The lexicon based approaches are known as semantic methods. These approaches rely on sentiment lexicons which are a collection of manually pre-compiled terms or words [Taboada et al., 2011]. The pre-compiled terms are then used to analyze the text to generate the sentiment classifications and their polarities. There are mainly two types of lexicon based sentiment analysis methods: Dictionary Based Approach, Corpus Based Approach.

The dictionary based approaches [Hu and Liu, 2004; Kim and Hovy, 2004] first manually collect a set of opinion words ‘seeds’ with their orientations (e.g., positive, negative) and relevant extents (e.g., high, moderate, low). Secondly, the seed set is expanded by searching synonyms and antonyms from well-defined corpora such as WordNet [Miller, 1995]. These new words are then added to the seed list. This process is repeated till no new words can be found. However, the dictionary based approaches do not perform well in different domains (e.g., crime, sports, politics) where context specific orientations are changed.

The corpus based approaches extend the solution strategy of dictionary driven techniques by adding the context specific orientations to the seed words [Abdulla et al., 2013]. Thus, it allows one to define corpus for multi-domains (e.g., sports) [Godbole et al., 2007]. The corpus based approaches also consider the syntactic language patterns that may occur along with the dictionary words. Although corpus approaches are applicable to multiple domains for sentiment classification, it is difficult to develop huge language corpus for variable nature of social information services. Other than the semantic methods, the corpus based approaches are often combined with statistical methods to train classifiers [Pak and Paroubek, 2010b; Melville et al., 2009].

2.3.2 Spatio-Temporal Sentiment Analysis

The advancements in the sensor and wireless communication technologies have enabled to accurately determine the location of users and objects [Gruteser and Grunwald, 2003]. In addition, the emergence of location sensing mobile devices and Web applications has encouraged users to share the information about their geographical locations with other users. There is a great interest in exploiting the users’ geographical locations for various applications [Rao and Minakakis, 2003].

Over time, many social information services such as Twitter, Facebook, Flicker, etc., have enabled social sensors to embed their geographical location information (e.g., longitude and latitude, point of interests) in their data by using geo-tagging or location check-ins options.
At the same time, social information services have the inherent feature to store the social sensor data based on the time of generation and sharing. The presence of these two features: time (i.e., temporal), and geographical location (i.e., spatial), in the social sensor data makes it more valuable. Consequently, it allows one to understand the social sensor activities (e.g., mobility) based on time and space [Cao et al., 2015]. The combination of sentiment analysis with spatio-temporal features enables the visualization and break down of social sensors' public opinion in more depth. For instance, sentiment of social sensors can be represented at multiple levels such as states, cities, etc. There are several efforts available that show the potentials of spatio-temporal sentiment analysis of social information services. For instance, a surveillance system is designed to monitor the Dengue epidemic in Brazil which collects the social sensor data to identify the disease effected regions [Gomide et al., 2011]. In [Mazumder et al., 2013], a visual map is developed that shows the radical sentiment of social sensors from different Indonesian states. In addition, social sensors’ sentiment and locations are analyzed together to develop a personalized location recommendation system [Yang et al., 2013]. Despite the provision of location sharing in many social information services, some social information services do not provide geo-tagging or check-ins facilities. In such circumstances, the geographical locations of social sensors are extracted by Named Entity Recognition (NER) methods [Malmasi and Dras, 2015]. The NER techniques work similar to corpus based approaches for location identification which require developing special corpus that contain the names of geographical locations. The NER techniques use time intensive semantic techniques of text parsing to retrieve geographical location names by matching the location specific corpus.

Overall, the domain of sentiment analysis is a broad area of research. However, both main types of sentiment analysis approaches and their subsequent methods are classified as ‘data-oriented’ techniques. These techniques have two main limitations. First, in order to perform sentiment analysis, one has to grasp multiple specialized skills such as statistical data modeling, machine learning and lexicon development. Secondly, both approaches require laborious and time-consuming tasks which include data collection, data preprocessing, manual data labelling, data model (i.e., algorithm) training, and model testing and validation.

2.4 Service Oriented Architecture (SOA)

There are several interpretations and definitions available of service oriented architecture (SOA) in industry and academia. For example, SOA is defined as a design strategy that
allows software to expose their functionality as services to other applications or services [Erl, 2008]. In contrast, The Open Group defines SOA as an architectural style that supports service orientation, whereas service orientation is a way of thinking in terms of services and service-based development, and the outcomes of services. Despite the variation in definitions and conceptualization of SOA, services are considered as pillars of SOA. Similar to the notion of SOA, the definition of ‘what is a service?’ is also presented differently by researchers and professionals. For instance, in [Sprott and Wilkes, 2004], a service is described as a component capable of performing a task. Similarly, in [Uleman, 2006], researchers defined a service as a well-defined, repeatable business task that can be performed by an application. In comparison, the organization for the Advancement of Structured Information Standards (OASIS) defines a service as a mechanism to enable access to one or more capabilities using a prescribed interface [MacKenzie et al., 2006]. In general, the design of SOA (see Figure 2.1) is realized in the following three components:

1. **Service Consumer** is a software or another service needing to perform a task. It finds the required service from a service registry and invokes the service function(s) execution at service provider’s end through service interface.

2. **Service Provider** is an entity that accepts and executes the service invocation requests from service consumers. Service provider publishes description of a service in the ser-
vice registry. The service description contains the information for service consumer to discover and access the service.

3. **Service Registry** is a repository which contains the descriptions of available services. It allows service consumers to query and locate the interfaces of required services.

SOA is a technology independent design pattern which does not rely on a specific set of frameworks, protocols and programming languages. It can be implemented via multiple technologies such as REST, Web services and Remote Procedure Calls (RPC) [Pautasso et al., 2008]. Each of the implementation strategies have their own features and underlying technical complexities. Most commonly used approach to realize SOA is through Web services. Web service based architecture is based on open source technologies such as eXtensible Markup Language (XML), Simple Object Access Protocol (SOAP), Universal Description, Discovery and Integration (UDDI), Web Services Description Language (WSDL) [Curbera et al., 2002]. SOAP is an XML-based protocol that uses existing transport protocols such as Hyper-Text Transfer Protocol (HTTP) to transport data. WSDL provides a computer readable description of a Web service. It includes the details of how to interact with the service such as data messages (i.e., input/output), communication protocol (e.g., HTTP), and service endpoints (i.e., Uniform Resource Locator (URL)). UDDI works as a service registry that contains the Web services descriptions. It enables service consumers to find service providers through a centralized repository of services.

### 2.4.1 Cloud Computing

The concept of service orientation computing (SOC) and SOA has different objectives. While SOA focuses on developing software applications in which software components are organized as services; SOC concentrates on utilizing services to support the rapid and low-cost development and easy composition of distributed applications in heterogeneous environments [Papazoglou et al., 2008]. In this regard, SOA is considered as a sub-component of SOC. The flexibility of SOC to encapsulate and integrate decentralized components (e.g., software) by using standardized protocols across the Internet has produced a new computing paradigm called cloud computing [Dillon et al., 2010].

Similar to the phenomenon of SOA and services, there are multiple definitions of cloud computing [Abdel-Basset et al., 2018]. For instance, the cloud computing paradigm is described as the Internet based on-demand assembling and delivery of resources (e.g., CPU, storage, software) as general utilities that can be rented and published by users [Zhang
et al., 2010]. Similarly, in [Foster et al., 2008], researchers defined cloud computing as a large-scale distributed computing paradigm driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet. On the other hand, the National Institute of Standards and Technology (NIST) defines the cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Following are the five essential characteristics [Mell and Grance, 2009] of cloud computing:

1. **On-Demand Self-Service** enables a cloud consumer to unilaterally and automatically provision resources as required without the human interaction with resource providers.

2. **Broad Network Access** provides access to cloud capabilities over the network via standard mechanisms by using thin or thick client platforms (e.g., mobile phones, laptops).

3. **Resource Pooling** allows providers to pool their resources to serve multiple consumers by using a multi-tenant model. This model can dynamically assign and reassign different physical and virtual resources according to consumer demands. Also, it gives a higher level of abstraction where consumers are not aware of the actual location of resources.

4. **Rapid Elasticity** resources are elastically provisioned and released and can be rapidly scaled outward and inward based on demand. For consumers, resources are usually presented in unlimited capacity that can be assembled in any quantity at any time.

5. **Measured Service** different types of resources are automatically controlled and optimized by pay-per-use or charge-per-use basis. The resource consumption can be transparently monitored, controlled, and reported by both the provider and consumer.

### 2.4.2 Cloud Services and Models

The resources in the cloud are shared with the clients as services. The cloud services are classified into three cloud service layers, and these services can be deployed in four models [Jadeja and Modi, 2012]. The details of cloud service layers and models are discussed below:
CHAPTER 2. BACKGROUND

Cloud Services

A cloud application is comprised of a set of combined hardware and software services. These services are built on the top of one another in a layered stack as shown in Figure 2.2. The cloud service layers are generally classified into three distinct layers.

1. **Software as a Service (SaaS)** provides on demand software applications to consumers over the Internet [Mathur and Nishchal, 2010]. SaaS eliminates the need to install the software on a client’s system. Thus, SaaS relieves the clients from the troubles of software installation, testing and maintenance. Some key SaaS examples are Google Apps, and Enterprise Resource Planning (ERP) systems such as Workday, SAP, etc.

2. **Platform as a Service (PaaS)** delivers software development frameworks including operating systems to consumers. While SaaS enables to utilize already made software solutions; PaaS allows consumers to develop and host their own cloud services and applications [Dillon et al., 2010]. Some of the PaaS examples are Google App Engine and Microsoft Windows Azure.

3. **Infrastructure as a Service (IaaS)** refers to the hardware or physical services (e.g., processing, storage, network bandwidth) provided to consumers [Bhardwaj et al., 2010].

![Figure 2.2: Cloud Service Layers](image-url)
In IaaS, consumers are independent to deploy software SaaS and PaaS application packages to meet their needs. Amazon EC2 and IBM Smart Cloud are some examples of IaaS providers.

**Cloud Models**

The cloud service as a solution can be implemented and deployed into four models. The four types of cloud models are defined as follows based on their deployment features:

1. **Public Cloud** sells its services to general public a pay-as-you-go manner [Fox et al., 2009]. It is owned, managed and operated by an organization or a combination of organizations such as businesses, academics or government organizations. The public cloud setup exists on the premises of the cloud provider. Some famous public cloud examples are Google App Engine and Amazon Elastic Cloud Compute (EC2).

2. **Private Cloud** is exclusively used by a single owner organization or business [Goyal, 2014]. Moreover, it is not available to the general public. Private clouds are usually managed by the owner organization or by a third party. Also, it may exist on or off premises. Microsoft private cloud and VMware vCloud suite private cloud, are some examples of private clouds.

3. **Community Cloud** is usually owned, managed and operated by one or more organizations [Ali, 2009]. This type of infrastructure is formulated as a shared resource pool where several organizations share similar requirements and goals. Combined email servers and Wikipedia are types of community clouds.

4. **Hybrid Cloud** is a inter-mixture of two or more cloud setups (e.g., public or private) [Sotomayor et al., 2009]. The hybrid cloud works in such way that the participating cloud setups remain independent entities with respect to data and applications, but allow to share computing resources for scalable performance. EMC corp and VMWare’s vCloud hybrid service are some examples of hybrid clouds.

**2.4.3 Service Composition**

A SOA based application is generally comprised of a set of services (i.e., atomic services) offering different functionality. Based on the requirements, atomic services are combined as a composite service. The process of developing a composite service is known as service
A service composition process is comprised of three steps: Service Discovery, Service Selection, Service Execution. In service discovery, a set of candidate services are retrieved from service registry to perform certain operations. However, it is possible that many services match the functional criteria. Thus, in service selection step, the retrieved services are ranked based on non-functional features. Finally, the selected services are executed as a composite service. There are two main approaches for service composition: Static Composition and Dynamic Composition.

**Static Composition**

In static composition, the aggregation of required services is defined at the time of application design. The composition of all necessary atomic services is determined, integrated and then the application is deployed. A static composition can be represented as an abstract model with a set of tasks which should be carried at the time of execution. In the abstract model, the tasks are fulfilled by atomic services which are statically bounded in a pre-determined order of execution. The static composition approaches are usually implemented via graph-based solutions. Several approaches are proposed based on combining Petri Nets and Business Process Execution Language (BPEL) for static service composition. Static service composition process mainly relies on syntactic description of services for discovery and selection process. As a result, the syntactic composition approach only focuses on finding the required services rather than the best available services. Also, static service composition is not a flexible mechanism to cater frequent changes in requirements.

**Dynamic Composition**

The dynamic service composition aims to cover the gaps of static service composition approaches. In dynamic service composition, the abstract model of tasks is defined and service selection process is completed automatically at run-time without the interaction of service requester. Thus, dynamic composition allows to modify, extend and adapt new requirements of service requester in a composition process. In comparison to syntactic methods used in static composition, the dynamic composition approaches imply semantic frameworks such as WSMO (Web Service Modeling Ontology) and OWL-S (Ontology Web Language - Service). Semantic approaches add meaningful
and computer readable descriptions for service discovery and optimal service selection. In addition, dynamic service composition enables to accommodate the non-functional or QoS requirements (e.g., response time, price) of service requester. Also, during service selection process, many functionally similar services are discovered. In such scenarios, QoS plays an important role in final service selection [Talantikite et al., 2009].

Dynamic service composition techniques are divided into two main categories: Workflow Based Approaches, Artificial Intelligence (AI) Planning Based Approaches [Moghaddam and Davis, 2014]. Workflow based approaches [El-Hadad et al., 2010; Bhiri et al., 2006] define an abstract composite service to an executable business process. First, the service requester specify their goals and preferences. These goals are then decomposed into abstract business processes with a set of tasks. Each task represents specific functions, and relevant control flows and data checks are imposed. Also, QoS requirements are specified for complete business process and individual tasks. Secondly, the required services are discovered and sorted based on QoS. Finally, the selected services are bind with the business process.

In AI planning based techniques [Gutierrez-Garcia and Sim, 2010; Kuzu and Cicelki, 2012], service composition process is visualized as a planning problem. A service planning problem is comprised of an initial state, a set of goal states, and actions with conditions to transit between different goal states. For service composition, an agent (i.e., planner) devices a plan to achieve the goal state (i.e., composite service) via actions (i.e., available atomic services). At the end, if a successful plan is generated, the selected services are executed in the planned order. In comparison to workflow based composition techniques, the AI based approaches do not require the predefined or abstract workflow models [Rao and Su, 2004].

To the best of our knowledge, in the context of social information service based sentiment analysis, there are no service composition techniques available. Rather, current sentiment analysis approaches and tools are generally developed by using previously elaborated traditional data-oriented methods.

2.5 Big Data

In recent times, the term ‘Big Data’ has been a center of attention in academia, industry and the media. Generally, the big data term is mainly used to describe the enormous amounts of datasets [Chen et al., 2014]. However, there is no single unified definition or terminology to describe big data [Ward and Barker, 2013]. For instance, big data is considered as intensive datasets which are unable to be captured, managed and processed by traditional
computational systems [Kaisler et al., 2013]. Similarly, NIST defines [Soumaya et al., 2017] big data as “Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies”. Despite the different orientations of big data, it is often characterized by its three features also know as ‘3Vs’ [Oussous et al., 2018; Bilal et al., 2016]. The 3Vs properties are defined as below:

1. **Volume** defines the size of the data that is measurable in units such as terabytes (TB), petabytes (PB), etc.

2. **Variety** show the heterogeneity of the data. The big data is comprised of different formats of data (e.g., text, images, sensor data) originating from various sources such mobile devices, wireless sensor networks, software logs and social information services.

3. **Velocity** depicts the frequency or speed of the data generation. It is measured in different time units such as millisecond, second, minute, hour, day, etc.

### 2.5.1 Big Social Data

Social information services have become an essential source of big data [Bello-Orgaz et al., 2016]. Evidently, a large chunk of big data is produced via social information services. Thus, the notion of ‘Big Social Data’ is specifically used to define the big data generated by social information services [Burgess and Bruns, 2012]. Moreover, big social data also adheres to the 3Vs properties of big data in terms of volume, variety and velocity. In contrast to big data which is produced and propagated through mixed sources such as physical devices, sensors, transaction logs, satellites, etc.; the prime distinction of big social data is that it is mainly yielded by social sensors and their activities. Thus, making the big social data and its potential applications more human centric. Big social data has been hailed as a key to analyze social behaviors by scholars, politicians, corporations, governments and journalists [Tufekci, 2014]. Due to the interest received from multiple areas, the domain specific analysis of big social data is a challenging task. Big social data is implicitly an interdisciplinary area of research which involves data mining, graph mining, machine learning, statistics, natural language processing, information retrieval, semantic Web and big data computing [Manovich, 2011; Cambria et al., 2013].
CHAPTER 2. BACKGROUND

2.5.2 Big Data and Cloud Services

Cloud computing is a powerful platform to perform large scale and complex computing operations [Hashem et al., 2015]. In comparison to traditional computing technologies, cloud computing offers parallel processing, virtualized resources and scalable data storage. As a result, cloud computing is an ideal solution for scalable big data processing and analysis [Bryant et al., 2015]. Over time, several solutions such as Hadoop, MapReduce, Parallel DBMS, etc., have been developed and implemented for big data processing and analysis [Agrawal et al., 2011]. However, these solutions are specifically design for data warehousing applications and large scale parallel database management systems (DBMS). In addition, the developed solution frameworks do not focus on big social data.

From the cloud service perspective, several efforts have explored the utility of cloud services for big data analysis applications. For instance, in [Zheng et al., 2013], an overview is presented that outlines the abstract functional architecture of big data as a service (BDaaS). It includes three service layers: Big Data Infrastructure-as-a-Service (BDIaaS), Big Data Platform-as-a-Service (BDPaaS), Big Data Analytics Software-as-a-Service (BDASaaS). The BDIaaS service layer leverages from IaaS in cloud and provides storage and processing services. The BDPaaS delivers powerful platforms to build big data analysis solutions. Similarly, BDASaaS provides applications and software to process large amounts of data and extracts useful information. Similarly, several architectures [Siriweera et al., 2015; Vu and Asal, 2015; Demirkan and Dursun, 2013] have been proposed to provide a set of abstract processing components for big data analysis applications. However, the main limitation of these architectures is that they only provide templates and high level conceptual models for big data processing and analysis. Furthermore, these frameworks do not specifically focus on big social data based sentiment analysis by using cloud services.

2.6 Chapter Summary

In this chapter, we have provided a brief summary of the important and relevant topics which provide background to our research. We presented the overview of social information services and their applications in various domains. We also studied the area of sentiment analysis, its traditional techniques and their limitations, and the combination of spatio-temporal characteristics with sentiment analysis of social information services. We then reviewed the SOA and its relationship with cloud computing paradigm. In addition, we briefly investigated the service composition approaches and their features. We surveyed
the concept of big data and its relationship with social information services. Moreover, we reviewed the association between cloud services and big data. We identified the limitations of existing service frameworks for big social data analysis. In Chapter 3, we investigate our first research question which develops a service composition framework that provides an alternative for traditional social information service based sentiment analysis approaches.
Chapter 3

Service Composition For Spatio-Temporal Sentiment Analysis

In this chapter, we propose a service-oriented framework ‘Sentiment Analysis as a Service’ (SAaaS) that extracts sentiment of social sensors from social information services, analyses and converts into useful information. We develop a service composition mechanism that integrates sentiment analysis services based on social information service classification. We evaluate the proposed framework with social information services based public health surveillance as a motivating scenario. In particular, we focus on the spatio-temporal properties to aggregate sentiment analysis results. The experiments are conducted on the real-world dataset, and results demonstrate the applicability of the proposed approach.

3.1 Introduction

The data generated by social sensors via social information services has two beneficial characteristics: 1) It contains the subjective information such as sentiments and opinions on different topics. 2) It holds the spatio-temporal information of social sensors. While the sentiment analysis or opinion mining mechanism helps to extract and learn human dynamics such as behaviors, patterns, attitudes and emotions from subjective information [Serrano-Guerrero et al., 2015]; the spatio-temporal information allows to gain insights of social sensor activities based on time and locations [Hwang et al., 2013]. Therefore, combining the spatio-temporal
properties with sentiment analysis allows to obtain better orientation of social sensors sentiments. In this chapter, we describe a service based architecture that uses services as a mean to collect data from social information services to extract sentiment based on time and location, and deliver it as a service.

In this chapter, we focus on two main contributions with regard to spatio-temporal sentiment analysis. First, the traditional sentiment analysis process is formulated by using service-oriented paradigm. This includes a service model that defines multiple sentiment analysis processes as service layers, and sub-process activities as composite services. Secondly, a service composition technique is devised for aggregating a set of services as a workflow for performing sentiment analysis. As a part of service composition, we introduce a classification model for social information services based on different properties of social sensor data. There are several types of social information services utilized by social sensors. These services provide different features and enforce several restrictions (e.g., text length limitations) on social sensors for the data sharing. Consequently, the social sensor data has diverse characteristics such as size, length and quality. These diverse features necessitate separate mechanisms for extracting the multiple types of information. Thus, the classification model enables adoption of most appropriate services for processing and analyzing the social sensor data of a particular social information service. In addition, we develop a semantic tag matching based model and an algorithm for sentiment analysis service composition. The approach extends the human centric tagging model [Dong et al., 2010] to add semantics for candidate component services such as data collection, noise removal and information analysis. Later, the semantic tag matching technique [Liu et al., 2011] is used to retrieve the relevant services based on the social information service classification model. To evaluate the proposed framework, we use flu surveillance as an example scenario. However, the framework is not restricted to disease surveillance applications, and it can be applied to different domains applicable to spatio-temporal based sentiment analysis.

The rest of the chapter is organized as follows. Section 3.2 describes the motivating scenario. Section 3.3 highlights the related work. Section 3.4 depicts the overview of framework and section 3.5 elaborates the details of system models and section 3.6 presents the semantic composition approach. Section 3.7 provides the details of experiment results and evaluation by using a real-world case study. Finally, the section 3.8 concludes the chapter.
CHAPTER 3. SERVICE COMPOSITION FOR SPATIO-TEMPORAL SENTIMENT ANALYSIS

3.2 Motivating Scenario

We illustrate public health surveillance as a motivating scenario. Let us assume that the Department of Health and Safety Services (DHSS) is interested in monitoring seasonal epidemics in Australia. DHSS wants to know about any recent ‘flu’ outbreaks in last week, within all states and territories. In addition, DHSS requires to classify the locations of affected people with high intensity of flu.

In this scenario, we assume that social sensors share their health issues on various social information services. Some social information services provide application programming interfaces (APIs) to search and collect the data. On the other hand, there are some social information services that do not provide such standardized mechanisms for data collection. In such case, third party services such as Web crawlers and scrapers can be used as an alternative to collect the data. Social information services allow social sensors to share their data as tweets, posts and comments with diverse features such as text length, data size, and noise levels. For instance, Twitter service enables social sensors to write a maximum 140 characters long text known as tweet\(^1\). Due to such limitations, social sensors use compact and misspelled words, abbreviations, emoticons and special characters in their textual data to express their opinion. In comparison, the Reddit blog service allows social sensors to write more than 1000 words in their blog posts and subsequent comments. With such flexibility, social sensors on Reddit service generally write their opinion in a formal writing style with proper English. As a result, these limitations and flexibility affect the quality or noise levels of social information services. Therefore, various types of noise removal services are required to filter out the noisy and irrelevant data from different types of social information services. Furthermore, many social information services do not offer the geo-tagging facility to social sensors for revealing their geographical locations. It is also possible that social sensors may not chose to expose their locations. In such cases, the geo-locations of social sensors are required to be extracted from the text, if available. Finally, due to dissimilar characteristics of data, single sentiment extraction service may not be able to analyze all types of social information services with equal accuracy. Thus, multiple sentiment analysis services are required for sentiment extraction.

In this scenario, by using the service composition framework for sentiment analysis (Figure 3.1), we first classify social information services by using different data properties. The

---

\(^1\)It should be noted that as of November 7, 2017, the character limit for writing tweets has been doubled for all languages except Japanese, Korean, and Chinese. However, throughout our thesis, our experiments and findings are based on the datasets which contain 140 character long tweets.
composition layer utilizes this classification for composing services required for sentiment analysis in four steps. First, the service composition layer selects the data collection services to gather the datasets regarding flu from various social information services. Secondly, the composition layer determines the suitable noise removal service(s) for individual datasets. Thirdly, geo-locations are abstracted from the datasets by location extraction services for the non geo-tagged data. In the last step, appropriate analysis services are selected for each dataset for sentiment extraction. Finally, the extracted analysis results containing the information of flu victims along with their sentiment polarity are presented to end users by using maps and charts for visualization.

3.3 Related Work

Various government departments and research organizations are committing resources in different initiatives for exploring novel and efficient ways of public health surveillance [Carroll et al., 2014]. For example, one such initiative in Australia is an online survey Website info.flutracking\(^2\). It crowd sources data from online participants regarding health issues such as flu. The participants are required to voluntarily fill-out weekly online surveys. On the other hand, there are several traditional disease surveillance methods available like interviews, paper based surveys and clinical reports [Lopez, 2007]. Despite the mode of data collection,
these methods generally rely on the volunteer participation of interested people. Moreover, these approaches are time consuming and lack efficiency. In comparison, our proposed framework provides an efficient approach that can be used as an alternative for traditional disease surveillance methods.

Social information services have become a free source for collecting the data shared by social sensor. One aspect of social information services is that social sensor propagate their data voluntarily without any hustle. Researchers and practitioners have investigated the role of social information services in the domain of health surveillance. For instance, in [Corley et al., 2010], text mining process is utilized on social information services and Web data to identify the relationship between flu and real patient reports. The research work [Parker et al., 2013] used tweets as opinions to detect the latest public health trends. Similarly, in [Gomide et al., 2011], the spatio-temporal abilities of social information services are highlighted. A dengue surveillance model is proposed to identify the locations of affected areas by using Twitter data. Although these approaches show the application of social information services in health surveillance, they rely on traditional text and data mining processes.

In contrast to traditional approaches for harvesting opinion form social information services, there are various online tools and platforms available for social information service monitoring and analysis [Batrinca and Treleaven, 2015]. These tools overcome the limitation of traditional opinion and sentiment analysis approaches by automating the data gathering and information extraction processes. However, online approaches have several limitations. Mostly, these tools focus on general-purpose social information service based search and analysis. This analysis may include basic results such as content specific statistics, social sensor clicks and actions (e.g., likes, dislikes) and trending public opinion. The end users of these tools are unable to perform complex opinion analysis searches [Maynard et al., 2012]. In addition, some tools are specifically designed for a single social information service. Consequently, when multiple social information services are involved in analysis, an end user will simultaneously require the use of different tools for a number of social information services. The usage of multiple tools is time consuming and provides discrepant views of the data [Wan and Paris, 2014].

Services provide a cheap solution to develop complex applications. In the essence of application development, services are characterized as black boxes with standardized interfaces where the underlying implementation is hidden from the end users [Mietzner et al., 2010]. The service paradigm delivers a powerful abstraction that hides the data specific information form end users and concentrate on how to use this data [Neiat et al., 2015]. An end user
visualizes services as functional blocks, and only the required blocks need to be composed as a workflow with respect to inputs and outputs. Typically, services are composed based on functional and QoS based requirements. While functional requirements focus on finding the service with required operations, the QoS requirements try to satisfy the non-functional demands such as price, response time, etc., while selecting a service [Medjahed et al., 2003]. There are several services (e.g., Web services and APIs) developed by academic bodies and commercial organizations for sentiment extraction and classification [Serrano-Guerrero et al., 2015]. These services are a viable solution option to develop sentiment analysis applications without indulging into technical complexities.

The potential applicability of service-orientation is recognized in the domain of sentiment analysis [Zaki et al., 2017]. However, to the best of our knowledge, there are no frameworks present in literature that uses any composition approach to aggregate services for social information service driven sentiment analysis. On the other hand, there are few research works [Lee et al., 2013; Musaev et al., 2014; Rosser et al., 2017] available which only focus on information extraction from multiple social information services and its fusion with various sources (e.g., sensors) for different analysis applications. However, these frameworks do not focus on composing services for abstracting spatio-temporal sentiment analysis. There are also several approaches proposed which consumes cloud as a platform for performing sentiment analysis. For instance, in [Piccialli et al., 2018], a service oriented system is proposed which leverages the cloud platform for real time data analysis. The proposed system combines the data from different sources including social information services and physical sensors, however, the sentiment analysis is performed using data processes rather via online services. Similarly, the framework [Jatwani et al., 2018] provides a cloud based sentiment analysis solution for analyzing social sensors behavior using Twitter service. However, it also does not compose services to perform sentiment analysis and its associated tasks. In addition, there are some frameworks proposed which outline the high level architectures or guidelines for utilizing services for sentiment analysis. For instance, researchers proposed a model [EL-Haddaoui et al., 2018] based on three abstract layer levels: Components Level, Orchestration Level, Services Level. The component level defines the sentiment analysis tasks for collecting, processing, analyzing and visualizing the data. The orchestration level outlines the resource allocation for sentiment analysis tasks. The services level defines the actual services performing the component level tasks. Similarly, in [Chen et al., 2019], researchers proposed a conceptual framework based on cloud services which can be realized to perform various analytics operations including sentiment analysis on online customer reviews.
In comparison to existing approaches, our proposed framework dynamically composes online services for collecting data from multiple social information services, processing and extracting sentiment. Moreover, provides the ability to perform sentiment analysis with respect to social sensors locations and time-lines.

3.4 The Framework Overview

The proposed framework (Figure 3.1) is divided into two logical categories: 1) System Models 2) Dynamic Service Composition Approach. System models define the baseline concepts and models that are consumed by the service composition layer for data extraction, processing and analysis. The dynamic service composition approach presents the underlying semantic tagging mechanism, service composition model and an algorithm for aggregating services.

The framework consists of four independent system components that perform the data collection, noise removal, geo-tagging and sentiment analysis. Each system component is conceptualized as an independent service layer with a set of similar services. We assume that services in each layer are provided by different service providers, and these service can be registered and retired in the respective layer at anytime. The components communicate through standard data messages in different formats such as Extensible Markup Language (XML), JavaScript Object Notation (JSON), and Comma-Separated Values (CSV). Figure 3.2 depicts the linear flow of the data between system components. The data collection component employs multiple services to obtain the raw data from various social information services. The noise removing component combines different data cleaning filters for preprocessing the raw data, and prepares the clean data which is ready for information extraction. The geo-tagging component extracts geo-locations information from non geo-tagged data. Finally,
the sentiment analysis component extracts the subjective information from the data such as sentiments, and emotions, and determines their polarity. The results are then presented in different visualization formats.

The composition framework integrates the system components and presents the combined resources by using the ‘as a service’ model. The ‘as a service’ model provides several benefits over traditional approaches, including separation of concerns, dynamic composition and abstraction. For instance, separation of concerns provides the benefits to formalize the proposed system in a set of loosely coupled layers comprised of several independent services. In each service layer, a service can be technology independent and hosted by a third party service provider. Moreover, a service can be developed from scratch and deployed as per requirements. The ‘as a service’ model allows services from different service layers to be dynamically compose on-demand based on functional and QoS properties. The abstraction hides the implementation details of services from end users. For example, end users are not aware that how the sentiments are extracted from text or how their polarity is computed.

In the next section, we formally define the three system models and present the details of their elements.

### 3.5 System Models

In this section, we describe the system models used in our framework. First, we define a formal service model that presents the functional and QoS features of services incorporated in each system component. Secondly, we classify social information services based on their generic data characteristics. Finally, we present a quality model that is used for filtering out noise from the raw data collected from social information services.

#### 3.5.1 Component and Composite Service Model

We use a top down approach for defining the service model. First, we define the SAaaS as a composite service. Later, we determine each system component and its component services.

- **Definition 1:** Sentiment Analysis as a Service (SAS) is defined as a composite service that is an end product of composition of component services. The SAS extracts sentiments from social information services for a given topic and delivers the analysis results to the service user. It is defined as a tuple of <\(ID, P, F_i, S - T, Q_i\)>, where
  - \(ID\) is the unique service identification number.
– $P$ is the sentiment analysis topic (e.g., flu) comprised of keywords for which component services are composed.

– $F_i$ is the set of various functions offered by the service such as sentiment aggregation, comparison and presentation.

– $S$-$T$ defines the spatio-temporal requirements to gather the data, where $S$ provides the requirements of targeted geographical area such as country or cities. $T$ represents the time bounds for which the analysis is required.

– $Q_i$ is a tuple of $<q_1, q_2, ..., q_n>$ non-functional features. Each $q_i$ denotes a unique QoS property such as price, response time, and service ratings.

• **Definition 2:** The data collection component is a service layer containing a set of data collection services which gather the data from various social information services. A data collection service $DS$ is a tuple of $<did, sk, pi, qi>$, where

  – $did$ is the unique identifier number of a data collection service.

  – $sk$ is an input parameter comprised of a set of search keywords or search terms to query social information services.

  – $pi$ is the targeted social information service from the data is to be collected based on temporal $T$ and; if possible by spatial $S$ parameters.

  – $qi$ is a set of QoS properties associated with the data collection service.

• **Definition 3:** The noise removal component is a service layer comprised of noise filters for removing different types of noise. A noise filter is visualized as a service $NS$. It is a tuple of $<nid, ds, tn, qi>$, where

  – $nid$ is the unique identifier number of a noise filtering service.

  – $ds$ is a set of data items $<d_1, d_2, ..., d_n>$ in a dataset, where a $di$ represents a data item such as tweet, comment or post.

  – $tn$ defines the type of noise removal filter and its functions.

  – $qi$ defines the non-functional features of noise removal service.

• **Definition 4:** The geo-tagging component is a service layer consisting of location extraction services. A geo-tagging service $GS$ extracts the social sensor locations from text, and assign the data items with location coordinate. $GS$ is defined as a tuple of $<gid, ds, Loc, qi>$, where
− gid is the unique identifier number of a geo-tagging service.
− ds is the dataset of a particular social information service for which locations are required to be extracted.
− Loc parameter determines the large data into fine-grained data segments based on location requirements $S$ by using top down approach. For instance, the data collected for Australia can be further divided into cities or towns.
− qi is a set of associated QoS properties of the geo-tagging service.

- **Definition 5:** The sentiment analysis component is a service layer constituting information analysis services. A sentiment analysis service extracts subjective information from the collected data. The subjective information can be classified into different types such as sentiment classification services which categorize a data item into: positive, negative, neutral, or emotion classification services which categorize a data item like happy, sad, angry, etc. A sentiment analysis service $AS$ is presented as a tuple of $<aid, op, ln, qi>$, where
  - $aid$ is the unique identifier number of a sentiment analysis service.
  - $op$ defines the multiple types of functions or operations like sentiment extraction and emotion extraction offered by a sentiment analysis service.
  - $ln$ determines the ability to perform the analysis on preferred human languages. Currently, we only focus on English language.
  - $qi$ determines the set of non-functional properties of a sentiment analysis service.

### 3.5.2 Social Information Service Classification Model

There are multiple types of social information services [Kaplan and Haenlein, 2010]. The usage of these services depends on the certain features which attract social sensors. As a result, social sensor data is affected by the social information service specific features. In addition, each social information service has different data generation patterns. For example, according to an online digital marketing company Zephoria Inc.\(^3\), every minute on Facebook 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded. Meanwhile, it is estimated that 500 million tweets are posted on Twitter everyday, 55 million photos are uploaded on Instagram per day, and 100 hours of videos are transferred

\(^3\)https://zephoria.com/
on YouTube per minute [Musaev et al., 2014]. Therefore, it is important to classify social information services based on their application usage and data features. The classification will allow selection of services which are best suited for information analysis process.

Social Information Service Types

We have categorized following four types of social information services based on their generic application usage and common features.

Content Community

Content Community services are mainly used for sharing media contents such as videos, audios and images. Social sensors upload their media contents on famous platforms such as YouTube, Instagram, and LiveLeak. The uploader can decide to make the content public or private. However, generally the content is available for public access. While other social sensors on these platforms express their sentiment and opinions in response to the shared contents.

Social Networking

Social networking services enable social sensors to create community networks or digital social networks. These networks can be utilized for personal, professional and leisure purposes. Social sensors on platforms like Facebook, Google Plus, and LinkedIn create personalize profiles, send invitations to friends, colleagues and family members to connect with their profiles. The data is shared via status posting, chatting messages and personal content sharing. Generally, the data shared by social sensors is not accessible for public.

Blogs

Blogs are simplest and oldest type of social information services. Social information services such as Reddit, Quora, and Rotten-Tomatoes are Web-sites that generally work as community pages and review sites. Blogs allow social sensors to upload contents and initiate online discussions relevant to the content. Blogs usually attracts social sensors who are interested in specific topics of interest and collaborative communities. The data available on blogs can be either private or public.
CHAPTER 3. SERVICE COMPOSITION FOR SPATIO-TEMPORAL SENTIMENT ANALYSIS

Table 3.1: Social Information Service Classification Model

<table>
<thead>
<tr>
<th>Service Features</th>
<th>Content Community</th>
<th>Social Networking</th>
<th>Blogs</th>
<th>Micro-Blogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Length</td>
<td>Medium</td>
<td>Medium-Long</td>
<td>Medium-Long</td>
<td>Short</td>
</tr>
<tr>
<td>Text Type</td>
<td>Informal</td>
<td>Informal</td>
<td>Formal</td>
<td>Informal</td>
</tr>
<tr>
<td>Data Volume</td>
<td>Stream</td>
<td>Stream</td>
<td>Non-Stream</td>
<td>Stream</td>
</tr>
</tbody>
</table>

Micro-Blogs

Micro-Blogs are the next generation derivatives of blogs. In comparison with blogs, social information services such as Twitter and Tumbler facilitate in sharing instant data with public. Social sensors share their data via small text messages (i.e., micro-posts) in broadcasting style. The data is generally available for public access.

Social Information Service Features

The social sensor data produced via above types of social information services have diverse features. We have generalized social information services with respect to three data features from sentiment analysis application perspective. Table 3.1 summarizes the social information classification model. Three features are explained as follows:

Text Length

Social information services often restrict social sensors for writing their textual data within a certain length. For instance, Twitter restricts to use maximum 140 characters, and Facebook allows its users to write posts with more than 60,000 characters. Meanwhile, social sensors posts can contain 100,000 characters on Google Plus. The available sentiment analysis services are developed to analyze text of different length. Thus, it is vital to classify social information services with their text length range for appropriate analysis service selection. We classify three types of text length with following ranges:

1. **Short**: A text length is classified as short, if it contains less than 500 characters.
2. **Medium**: In this type, the text length of a data item is more than 500, but less than 5000 characters.
3. **Long**: The text length is classified as long, if it contains more than 5000 characters.
Text Type

The natural human languages are ambiguous in nature. The writing style of a language is affected by imposed restrictions of social information services, and social sensor driven factors such as age, gender, language competency, and popular trends. We classify two text types used by social sensors: Formal, Informal. In informal text type, the text writing of social sensors is widely influenced by the Internet language. This includes various trending abbreviations, slang words, emoticons and special characters to express opinion [Thelwall et al., 2010]. In formal text type, social sensors use proper language skills and avoid using Internet language in scenarios such as writing product reviews, answering online questions, etc. This classification enables one to find suitable analysis services based on social information service specific text type.

Data Volume

Social information services continuously generate large data for emergency and daily life events. However, the volume and generation of the data relies on the social information service type and the number of participating social sensors. We categorized two types of data volume: Stream Data, Non-stream Data. Stream data contains a large number of contributing social sensors and the data is generated within limited amount of time [Kavanaugh et al., 2012]. For instance, if an incident occurs such as a natural disaster, social sensors overwhelm social information services with their opinions within hours. In contrast, non-stream data incorporates a limited number of social sensors who share data not related to any sudden event, and the amount of the data generated is not excessive. The data volume feature of a social information service helps to decide selection of analysis services based on their ability to efficiently handle datasets of different sizes.

3.5.3 Social Information Service Quality Model

The raw data collected from social information services have different types of noise [Zeng et al., 2010]. Removing irrelevant and noisy data is a challenge. The data noise raises the overall processing time of sentiment analysis. If noise is not removed, it may change the outcomes and hinders the reliability of sentiment analysis results. We devise a social information service quality model that removes the different types of noise from collected raw data. The quality model acts as a set of noise filters. During service composition process, it is possible that these filters are applied for a social information service in different sequence and combination
for the data cleaning. The quality model is comprised of following three filters:

1. **Data Overload:** The data obtained from social information services often contains unnecessary data such as embedded URLs, special characters, and emoticons. The data overload \(DO\) filter discards the unnecessary data. The \(DO\) is defined as follows:

\[
DO = \frac{UD}{RD}
\]  

(3.1)

Where \(UD\) is set of data items successfully removed of unnecessary elements, and \(RD\) is the total number of data items in the retrieved raw dataset.

2. **Data Relevancy:** The data obtained from social information services often contains irrelevant contextual data. For instance, in this tweet “I liked a @YouTube video youtu.be/LUID0jSh2Ic?a Saturday Night Fever (Bee Gees, You Should be Dancing)” is collected by using the keyword ‘fever’. However, it is not relevant to the context of fever as a disease. The data relevancy \(DR\) filter removes the contextual irrelevant data items from the retrieved data. The \(DR\) is defined as below:

\[
DR = \frac{RLD \cap RD}{RD}
\]  

(3.2)

Where \(RLD\) is contextually relevant data items, and \(RD\) is the total number of data items in the retrieved raw dataset.

3. **Data Corruption:** The data items in a dataset often contain corrupted data elements such as misspelled words or intentional typing errors. For instance, the following tweet contains an intentional typos: “I hate this flu soooooooooo much........... :(. The data corruption \(DC\) removes such discrepancies and errors. The \(DC\) is defined as follows:

\[
DC = \frac{DE}{RD}
\]  

(3.3)

Where \(DE\) is the corrected data items, and \(RD\) is the total number of data items in the retrieved raw dataset.

In next section, we present our service composition approach which utilizes the social information service types and features defined in the classification model (see Table 3.1).
3.6 Dynamic Service Composition Approach

The service composition layer in our proposed framework is responsible for composing services from each service layer. The composition method employs the social information service classification model as pivot for service selection and composition. The composition approach assumes that there are a set of functionally similar services available for each service layer provided by multiple service providers. The composition approach uses a semantic matching methodology for service composition. First, we extend and apply an existing human-centered concept tagging model [Dong et al., 2010] that labels candidate services with semantic tags. Second, we devise an algorithm which retrieves and aggregate the services by matching semantic tags from each service layer as a workflow. The remainder of the section provide the details of the approach.

3.6.1 Human Centric Tag Based Semantic Model

The service providers publish their services in a registry infrastructure such as UDDI by using service description formats (e.g., WSDL). The service consumers can look-up the service from the registry by using the service description. In our tagging model, we assume the similar infrastructure for service registration. However, in our proposed model, service providers cannot directly register their services. First, a service provider sends a request to human user (i.e., registry admin) for service registration. The registry admin receives the requests and assigns the relevant semantic tags to the service. Then, it is added in the framework’s service registry. Later, the service composition layer can retrieve the service information from the registry. Similarly, the registry admin can also develop or find a service, and register it by using the same process.

The registry admin defines the higher abstraction of a service known as service concept. A service concept is a combination of service properties represented by semantic annotations. Semantic annotations are simple meaningful terms. These annotations are added as tags with each candidate service at the time of registration. In this semantic model, social information service types and features are used as service concepts and their corresponding values are considered as service properties. For instance, the admin registers a tweet collection service, Twitter Streaming API, as follows. First the requested candidate service is assigned a service layer of Data Collection. As a next step, it is assigned a concept of Social Information Service Type with the annotation Micro-Blogging. It is possible that a service can be assigned multiple concepts but it will only have one service layer. For example, a data collection service
CHAPTER 3. SERVICE COMPOSITION FOR SPATIO-TEMPORAL SENTIMENT ANALYSIS

may offer to gather data from multiple social information services. The service concepts are used by composition layer for service discovery from each service layer. The semantic tagging model is formally presented in three following elements.

- **Candidate Services** is a set \( S = \{s_1, s_2, ..., s_M\} \) of \( M \) number of candidate services.

- **Service Concepts** represent the set \( C = \{c_1, c_2, ..., c_N\} \) of \( N \) number of service abstractions as service layers.

- **Service Properties** is a set \( P = \{p_1, p_2, ..., p_K\} \) of \( K \) number of semantic annotations belonging to a service concept.

The taxonomy of the semantic model for a candidate service \( s_i \) is defined as follows:

\[
s_i = \begin{cases} 
(s_i, c_i, p_K), & \text{where } i, K > 0 \\
p_K \in c_i \\
 s_i(p_K) = \{p_1, p_2, ..., p_K\}
\end{cases}
\] (3.4)

The above taxonomy shows that a candidate service \( s_i \) is tagged with a concept \( c_i \) and corresponding annotation \( p_k \) as tuple \( < s_i, c_i, p_k > \).

### 3.6.2 Social Information Service Classification Based Concept Tagging

We utilize the social information service classification model as a blue print for the semantic tagging of candidate services. The candidate services in service registry are clustered by using concept tags, whereas a cluster \( CT \) constitutes a service layer. By using the above semantic taxonomy, the semantic tagging of candidate services for four layers in service registry is illustrated as below:

- **Data Collection**: A data collection service \( DS \) is tagged as a tuple of \( < DS, Data Collection, Social Information Service Type > \).

- **Noise Removal**: A noise removal service \( NS \) is tagged as a tuple of \( < NS, Noise Removal, Text Type > \).

- **Geo-Tagging**: A geo-tagging service \( GS \) is tagged as a tuple of \( < GS, Geo-Tagging, Text Length, Data Volume > \). Although many social information services do not provide geo-tagged data, a non-geo tagged dataset is processed with the geo-tagging service.
CHAPTER 3. SERVICE COMPOSITION FOR SPATIO-TEMPORAL SENTIMENT ANALYSIS

- **Sentiment Analysis**: A sentiment analysis service $\text{AS}$ is tagged with three concepts: Text Type, Text Length and Data Volume. It is represented as a tuple of $< \text{AS}, \text{Sentiment Analysis, Text Type, Text Length, Data Volume}>$.

### 3.6.3 Classification-Driven Service Composition

For the service composition process, we utilize a sequential composition [Sun and Jing, 2012] order for aggregating services as the data flows between the service layers in a defined forward sequence. For example, the data collection service layer inputs the collected data to noise removal service layer. The noise removal service layer transfers the filtered data to geo-tagging service layer which finally sends the data to sentiment analysis service layer. The composition can be illustrated by using a DAG (Directed Acyclic Graph). A DAG is comprised of three nodes: **Initial State (IS)**, **Transitional State (TS)**, **Final State (FS)**. In terms of composition process, the IS is the starting node. At IS, a targeted social information service and relevant data capturing parameters are determined. TS captures the effects of four service layers: Data Collection, Noise Removing, Geo-Tagging, Sentiment Analysis, as four transition states through which data flows in a linear sequence. FS establishes the final output comprising of a set of services in the composition.

The service discovery and selection occurs during the transition to TS states. The service composition algorithm determines the appropriate service for each state and then transits to the next state. A service $s_i$ for TS is deemed composable, if it satisfies two rules. First, candidate service in the registry must be compatible with the current state TS. Secondly, the service property tags or annotations must belong to the social information service classification table $\text{DT}$. The composition rules are stated as follows:

\[
\text{Composable}(s_i) = \begin{cases} 
(s_i \in CT), \text{ where } CT = TS \\
 s_i(p_K) \in DT 
\end{cases} \tag{3.5}
\]

To illustrate the service composition process (see Figure 3.3), we use the motivating scenario as an example. Let us assume that end user selects Facebook as a social information service with spatio-temporal parameters for flu surveillance. The composition layer initiates the composition process and assigns the Facebook and parameters as an input to IS. For the data collection state TS, the composition layer finds a composable data collection service $\text{DS}$ for Facebook from the registry in the data collection cluster $< \text{CT: DataCollection}>$. The composition layer abstracts the service property tag of ‘Social Information Service Type’ and look-
up classification table $DT$ for the associated properties under the ‘Social Networking’ column. Then, the composition layer transits to the next state of $NoiseRemoval$. At this state, the composer selects a noise removal service from noise removal cluster $<CT: NoiseRemoval>$ that is applicable for text type property with ‘Informal’ value. Similarly, the composition layer finds the appropriate services for next two transit states of geo-tagging and sentiment analysis from clusters: $<CT: GeoTagging>$, $<CT: SentimentAnalysis>$, respectively, by using the social information service type tag. In complete composition process, the semantic tags of candidate services and classification table are used as semantic validation conditions. Algorithm 1 formally defines the composition process.

On the other hand, it is possible that during service discovery process, more than one candidate services matches the relevant properties. In such scenario, the service composition layer utilizes the QoS of a candidate service as final service selection criteria. For instance, if an end user has price and accuracy constraints then each candidate service is selected which best suited the non-functional requirements. The QoS based service composition is a well-established area of research in service computing. Over time, there are several methods and approaches proposed for specifically QoS driven service selection and composition [Alrifai and Risse, 2009]. We assume an existing QoS mechanism [Zeng et al., 2004] based on Multiple Criteria Decision Making (MCDM) for selecting a candidate service from each service layer. The following function calculates the score of a candidate service:

$$Score(s_i) = \sum_{j=1}^{n} (V_{ij} \ast W_j)$$  \hspace{1cm} (3.6)

For each service layer, a row $V_{ij}$ corresponds to a candidate service $s_i$, whereas each column corresponds to a quality attribute $j$ with normalized values of the service. A user provides a
Algorithm 1 Classification-Driven Based Composition

**Input:** Social Information Services List \(< SIS >\), Classification Table (DT)

**Output:** Service Composition

1. **for** Each service in \(< SIS >\) **do**
2. Initialize DAG(IS)
3. Get Composable Service \((DS:SIS \rightarrow CT:Data Collection)\)
4. Sort By QoS()
5. Look-up Semantic Tag Properties \(T \leftarrow DT\)
6. Transit()
7. Get Composable Service \((NS:SIS:T \rightarrow CT:Noise Removal)\)
8. Sort By QoS()
9. Transit()
10. Get Composable Service \((GS:SIS:T \rightarrow CT:Geo Tagging)\)
11. Sort By QoS()
12. Transit()
13. Get Composable Service \((SA:SIS:T \rightarrow CT:Sentiment Analysis)\)
14. Sort By QoS()
15. Transit()
16. Finalize DAG(FS) \(\rightarrow\) Service Composition List \(< DAG >\)
17. return List \(< DAG >\)

list of \(W_j\) which presents the weight of criterion for a quality attribute, where \(W_j \in [0,1]\) and \(\sum_{j=1}^{n} W_j = 1\). We assume that the SAaaS user provides the QoS preferences weights \(W_j\) which are mapped to each service layer. A candidate services with highest weighted score is selected from each service layer.

### 3.7 Experiments and Evaluation of Proposed Framework

We have conducted experiments to evaluate the performance and application of our proposed service composition approach. We divide our experiments into two parts. In first set of experiment, we present the applicability of our framework by using online services for sentiment analysis on real-world data. We use public health surveillance as an implementation use case scenario. Currently, we only focus on demonstrating the conceptual idea of performing sentiment analysis by using the service composition. In Chapter 5, we compare the accuracy of service composition driven sentiment analysis approach with traditional sentiment analysis approaches. For the second experiment part, to the best of our knowledge, there is no publicly available test dataset to measure the performance (i.e., execution time) of service composition for social information services based sentiment analysis. Thus, to evaluate the execution performance of service composition algorithm, we employ synthetic dataset of candidate services for evaluating the execution performance of service composition
CHAPTER 3. SERVICE COMPOSITION FOR SPATIO-TEMPORAL SENTIMENT ANALYSIS

for sentiment analysis. In the next subsection, we discuss our experiment configuration. In the later subsections, we present our experimental results.

3.7.1 Experimental Setup

As a proof of concept, we developed a system prototype comprised of various services. In our prototype, we have collected data from various social information services for monitoring seasonal epidemic of ‘Hay Fever’ and identified the locations of affected people along with their sentiment. For the performance evaluation, we focus on evaluating our composition approach by simulating the synthetic services. For each service layer, we simulate a set of candidate services and compute the service composition time where the candidate services are selected randomly. The experiments for service composition process are repeated 5 times and the average time is computed.

The prototype system is developed by using Microsoft Visual Studio 2015 .Net framework with ASP.Net/C#. The performance evaluation simulation of service composition is developed using python. The experiments are conducted on a 3.40 GHz Core i7 processor and 8 GB RAM by using Windows 7 as operating system.

3.7.2 Health Surveillance: Case Study

We use Hay Fever (allergic rhinitis) epidemic as our case study. On the evening of 21st November 2016 followed by a thunderstorm event, several hospital emergency departments in Victoria, Australia, observed a sharp increase in patients exhibiting with respiratory symptoms. Eventually on 22nd November, the Chief Health Board Advisory issued a health alert and declared a medical emergency. The emergency event overwhelmed the local ambulance services and hospitals, caused at least eight deaths. Epidemic caused by thunderstorm asthma events are rare and hard to predict. However, usually they can occur during October to mid-December. Such abrupt events can affect people who are suffering with Hay Fever and different respiratory conditions. Hence, a timely planned vaccination campaign by health departments for the vulnerable and affected people can help to prevent fatal consequences.

For prototype demonstration, we have collected the data from various social information services which is generated by social sensors before the declaration of health emergency. We identify the locations and sentiment of social sensors suffering with Hay Fever symptoms.

CHAPTER 3. SERVICE COMPOSITION FOR SPATIO-TEMPORAL SENTIMENT ANALYSIS

Table 3.2: Noise Filtering Results

<table>
<thead>
<tr>
<th></th>
<th>Data Overload</th>
<th>Data Relevancy</th>
<th>Data Corruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>31.6%</td>
<td>–</td>
<td>21.3%</td>
</tr>
<tr>
<td>SNR</td>
<td>–</td>
<td>12.8</td>
<td>–</td>
</tr>
</tbody>
</table>

Data Collection

We have collected the social sensor data from four social information services: Twitter, Reddit, Instagram and news blogs. The data from these social information services was collected using different mechanisms. For instance, we collected data from Twitter by its platform provided *Twitter Search API*. It provides the ability to search and gather tweets based on keywords. We used four search keywords: ‘hayfever’, ‘flu’, ‘allergy’ and ‘asthma’ for the data collection. On the other hand, we used third party online Web crawler and scraper services to collect data from other three social information services. Our dataset contains 525 reviews comprising of comments, blog posts and tweets. The data was collected during the months of October and November. All of the collected data is stored into CSV (Comma Separated Values) file formats.

Noise Removal

The social information service quality model is used as filtering mechanism for noise removal. Each noise removing filter is visualized as a noise removing service. To evaluate the performance of the filtering services, we use *Accuracy* and *Signal-To-Noise Ratio (SNR)* as evaluation measures. *Accuracy* is used to evaluate the *Data Overload (DO)* and *Data Corruption (DC)*, while *SNR* is utilized to assess *Data Relevancy (DR)*. The *Accuracy* and *Signal-To-Noise Ratio (SNR)* are defined as follows:

\[
\text{Accuracy} = \frac{\text{AccurateReviews}}{\text{TotalReviews}}
\]  \hspace{1cm} (3.7)

where \(\text{TotalReviews}\) is the number of all reviews in the dataset and \(\text{AccurateReviews}\) shows the number of reviews processed successfully.

\[
\text{SNR} = \frac{\text{Signal}}{\text{Noise} = \text{Total} - \text{Signal}}
\]  \hspace{1cm} (3.8)

where \(\text{Signal}\) is the number of reviews that are relevant and \(\text{Total}\) is the total number of reviews in the dataset. Table 3.2 shows the results of noise removal filters. For the
data overload, we used several customized regular expressions to discard the unnecessary
data such as special characters and embedded URLs. The data collected from Twitter and
Instagram mostly contained the unnecessary data. 31.6% data items were eliminated of
unnecessary data. It is important to highlight here that to gain maximum accuracy for data
overload filter, the regular expressions can be designed as per dataset. Next, we used a set
of stop words such as ‘Justin Beiber Fever, Beiber Fever, Songs, Videos, Dance’ to exclude
the irrelevant data items. Stop words are mutually exclusive terms or phrases which cannot
occur in a data item along with search terms. Due to smaller dataset, the SNR of 12.8 is
higher than expected. However, SNR may fluctuate if the search and stop keywords are not
selected appropriately. Finally, we used a dictionary based approach for the data correction
in the third filter. The performance of the data correction filter is lower than expected. Only
21.3% data items were rectified. In addition, we observed that using data correction filter
is a time and resource intensive task which may not be scalable to large datasets.

Geo-Tagging

Our dataset is gathered from four distinct social information service, only Twitter has the
geo-tagged data that reveals the social sensors location. The location information for the
non-geo tagged data was extracted from each data item. We used Named Entity Recognizer
(NER)\(^5\) for location extraction. It extracts geo-location names such as cities and towns
from data items by text parsing. After parsing the text and extracting location names, we
allotted the geo-coordinates to each data item by using a gazetteer. The public gazetteer\(^6\)
is developed by the Australian geo-science department. It contains a database which has
the information of more than 370,000 geographical places across Australia and its external
territories mapped with their longitudes and latitudes. We also evaluated the effectiveness
of geo-tagging by using SNR. The SNR value of location detection service is 4.8 for total
data. The data items without location information are discarded. We observed that social
sensor preferred to use the city names rather suburbs or states. For the sake of simplicity,
we grouped the data by two locations: Melbourne and Sydney.

\(^5\)http://nlp.stanford.edu/software/CRF-NER.shtml
\(^6\)http://www.ga.gov.au/placename
Sentiment Analysis

For sentiment classification, the dataset is divided into three subsets based on their applicable social information service type. The Twitter service is classified as micro-blogging, while Instagram is categorized as content community. Other two services are classified as Blogs. Based on the social information service classification and associated properties, we selected two sentiment analysis services: Alchemy API and SentiStrength [Thelwall et al., 2010]. The Alchemy API is a natural language processing service commercially available for sentiment extraction. This service can analyze both formal/informal types of text with medium to long length text, and it is suitable for analyzing blog reviews and news articles. Thus, we employ it on the data collected from Reddit and News Blogs. On the other hand, SentiStrength is a sentiment analysis service which is specifically designed to analyze informal and short length text. Hence, we utilized SentiStrength service for Twitter and Instagram data.

Prototype Results

After completing noise removal and geo-tagging, 121 data items were discarded from the dataset. The remaining 404 data items are grouped based on their social sensors location. Out of 404 data items, 284 reviews are generated from Melbourne, Victoria, and 120 reviews from Sydney, New South Wales. Figure 3.4 demonstrates the detected location results of social sensors suffering with Hay Fever on the Google map. As a result of sentiment classification, out of 284 Melbourne based reviews: 175 classified as negative, 54 positive, and 55 neutral. In comparison, among 120 Sydney based reviews: 58 classified as negative, 27 positive and 35 neutral. We present sentiment classification with three color schemes. The Red, Green and Yellow colors depict negative, positive and neutral sentiment classifications respectively. Figure 3.5 summarizes the sentiment classification results for both cities.

The outcome of the prototype results can be concluded in two dimensions. First, the collected data is obtained before declaring the emergency by the health department. Thus, it proves that a large number of social sensors were already suffering from hay fever symptoms. The social sensor data could have been helped the health departments of Australia to take precautionary actions such as awareness efforts and vaccination campaigns to minimize the fatal consequences of the epidemic. Secondly, the experiment results show the applicability of our approach which uses online available services for social information service based spatio-temporal sentiment analysis.

CHAPTER 3. SERVICE COMPOSITION FOR SPATIO-TEMPORAL SENTIMENT ANALYSIS

Figure 3.4: Hayfever Surveillance Map: Detected Social Sensors With Hayfever

Figure 3.5: Aggregated Sentiment Analysis Results of Social Sensors
3.7.3 Evaluation of Service Composition Approach

In the second set of experiments, we evaluate the performance of our proposed semantic service composition approach. Our main aim is to study the scalability of the composition algorithm by simulating the execution time of service composition process. First, we generate synthetic candidate services for four service layers: Data Collection, Noise Removal, Geo-Tagging, and Sentiment Analysis. A candidate service is produced as a tuple based on its definition features in respective service layer. Each candidate service is then assigned random semantic classification properties as per semantic tagging model. In this experiment, we do not consider the changing end user’s QoS preferences. However, to keep the maximum variability in our experiment, we randomly assigned an aggregated QoS score to a candidate service between a normalize score range of $[0, 9]$. Finally, we executed the service composition algorithm which selects a relevant candidate service from each service layer by matching semantic tags, and ranks the services by highest to lowest QoS values. We gradually increase the number of candidate services in each service layer from 2000 to 10,000 with the increment of 2000. Each increment of candidate services is executed by the composition algorithm 10 times and the average composition time is calculated. Figure 3.6 shows the execution time of composition algorithm with respect to the number of candidate services per service layer. Due to the nature of linear composition process where services are selected one by one from
four service layers, the increase in service selection time also increases in a scalable and consistent order.

3.8 Chapter Summary

This chapter presented a service-oriented approach that can be used as an alternative for traditional sentiment analysis process. We developed a service framework approach that collects data from multiple social information services, analyses and transforms into spatio-temporal sentiment analysis. We devised a classification model of social information services based on their generic data properties. We developed a semantic service composition technique that composes appropriate services for sentiment analysis based on social information service classification. We also introduced a quality model to remove noise from social information services. To demonstrate the performance of our proposed approach, we have conducted several types of experiments. We used a health surveillance scenario that applied sentiment analysis on real-world data to identify the locations of Hay Fever victims. In addition, we evaluated the scalability performance of our semantic service composition approach. The results show the applicability of our proposed approach.

Limitations and Future Work: Currently, our service composition approach is mainly driven by the static classification properties of social information services. In contrast, the social information service features are highly dynamic. Thus, an on-demand mechanism is required which can dynamically assess the relevant features and compose services for sentiment analysis. In addition, we have conducted our experiments with limited data, whereas social information service produce large amounts of data. In subsequent chapters, we aim to enhance our current service composition framework that is capable of dynamically evaluating changing features and composing services for large streams of social information services data. We also plan to compare the accuracy of our sentiment analysis approach with existing techniques.
Chapter 4

Social Information Services: The Service-Orientation of Social Media

In Chapter 3, we developed a service composition framework that can be used as an alternative for the conventional approaches to perform social information services based sentiment analysis. Initially, we used the notion of social information services to depict social media platforms, and defined a classification model based on their generic features. Our composition framework used the classification model as a blueprint to orchestrate the service selection for spatio-temporal sentiment analysis. However, the classification model remains a static mechanism that dictates the service composition process. In contrast, the cyberspace of social information services is highly dynamic and complex in terms of their features. Moreover, different domains or topic of interests such as health, entertainment, and politics may affect the data characteristics of social information services. Therefore, in order to best utilize a social information service in composition scenarios requiring different processing and analysis service; it is vital to take into account its dynamic features at service level.

In this chapter, we investigate how to effectively interpret the diversity of individual social information services, and how these services can be discovered, manipulated and accessed for different composition scenarios in a Web accessible environment. We establish the groundwork to extend our composition framework that will enable to dynamically discover and assess the features of different sets of social information services, and compose appropriate sentiment analysis services without relying on social information service classification model.
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

4.1 Introduction

Social information services attract social sensors from all over the world. Initially, social information services were recognized as sites that would allow the connectivity between distant social sensors for online communication. However, over time, social information services have become online platforms which not only enable online communication, but also data sharing platforms for social sensors with similar interests. A topic of interest can be related to politics, a new movie release, or a recent sport event. For instance, a social sensor shares a dance video of an actor on YouTube service, and the interested social sensors share their sentiment and opinion in response. This phenomenon has enabled social information services to be visualized as data sources which contain the data pertaining to different topic of interests (i.e., domains). Consequently, domain specific applications such as business, product design, disaster management, and health surveillance can collect the relevant data and analyze it for various purposes.

Social information services provide many overlapping functions such as chatting, image sharing, and video streaming to social sensors. Despite these functions serving the common purpose of data sharing, social information services differ in terms of social sensor participation, application usage, platform specific restrictions and their data sharing mechanisms [Dai et al., 2007]. As a natural consequence, social information services produce data with varying features like length, noise, and frequency. Regardless of differences in features, current analysis tools consider social information services as a homogeneous data source with similar characteristics. In the previous chapter, we established a baseline classification model to generalize the different types and features of social information services. Based on the classification model, the service composition layer composes sentiment analysis services for a social information service by assuming that it has a certain set of features. However, the composition framework is unable to dynamically assess the diverse and changing features of individual social information services prior to service composition. Hence, in order to efficiently compose sentiment analysis services based on changing features, it is essential to understand the diversity of social information services and their heterogeneous data features.

In this chapter, we present a novel mechanism that interprets the diverse and dynamic features of social information services. We apply the notion of ‘Service-Orientation’ for modeling and analyzing social information services. The service-orientation of social information services will provide two key benefits. First, the functional and QoS features of social information services can be defined and analyzed. Secondly, it would enable to access and
manipulate a social information service similar to a traditional Web/cloud enabled service based on its features. As a result, the service composition layer would be able to compose appropriate sentiment analysis services without relying on the predefined properties of social information service classification model. The contributions of this chapter are summarized below:

- We define an abstract service model to define social information services as cloud/Web services. We design two models to present the functional and non-functional properties of social information services.

- We collect the data from three social information services: Facebook, Twitter, YouTube, for three different domains: politics, entertainment, health. We analyze and quantify results based on developed functional and non-functional models.

- We present a Web accessible service model that formalizes social information services by using Ontology Web Language for Service (OWL-S). The formal service model provides a blueprint that enables the automatic discovery, selection and composition of social information services.

The rest of the chapter is organized as follows: Section 4.2 illustrates the motivating scenario. Section 4.3 highlights the related work. Section 4.4 describes the details of the design and analysis methodology. Sections 4.5, 4.6 and 4.7 provide the results of our methodology. Section 4.8 provides the summary of our findings and chapter.

4.2 Motivating Scenario

In order to understand the significance of service-orientation of social information services, we extend the previous social information services based sentiment analysis as our motivating scenario. Bella is a social information service analyst working for the Department of Health Services (DHS) and the Department of Social Services (DSS). Let us assume that both of the departments require Bella to develop two separate sentiment analysis systems. The DSS requires to analyze citizen sentiment on the government’s new policies to understand the general public views on social and political issues. In contrast, the DHS needs a system that can detect epidemic outbreaks (e.g., flu) and their severity within the country. Due to time and cost constraints, both departments do not want to develop their systems from scratch.

A traditional data-oriented approach for sentiment analysis constitutes three main steps: data collection, data preprocessing (e.g., irrelevant data filtering), data analysis (i.e., infor-
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

By using the conventional method, Bella has to collect raw data from various social information services by using several tools or service. Secondly, she requires to remove the noise from the raw data. Thirdly, the cleansed data is divided into two datasets: training, validation. Based on the analysis requirements, a machine learning algorithm is trained with the training dataset and tested with validation dataset. The training and validation process is repeated, until the desired accuracy of the algorithm is achieved. Above approach needs several manual and semi-automatic tasks including manual dataset labeling, algorithm training, and validation for information extraction [Tinoco et al., 2017]. Bella finds that the above approach is time consuming and lacks the automation.

To overcome the above limitations, Bella utilizes the service-oriented strategy to develop the sentiment analysis systems by using online services. Bella defines two diverse strategies for both departments as follows. For DSS, Bella first selects a set of social information services for the data collection that are suitable for social discussions. Secondly, she chooses services for the data preprocessing. Finally, she needs the sentiment analysis services that are specialized for extracting information from politics and social discussions. It is possible that a single tool or service for preprocessing and sentiment analysis is not applicable for all social information services (e.g., Twitter, YouTube), as selected social information services have different functional and non-functional properties. Therefore, Bella selects multiple preprocessing and sentiment analysis services for each social information service. For DHS, Bella adopts a similar strategy with following changes. She selects a set of social information services suitable for health information sharing. However, this time as part of requirements, Bella only selects social information services which have the information of social sensor ge-
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

...ographical locations. Alternatively, she chooses services that can extract the geographical location information by parsing the social sensor data. Afterwards, similar to her first strategy, she selects appropriate services for the data preprocessing, and specialized sentiment analysis services to extract the health information. Finally, the best available services are aggregated as workflows (see Figure 4.1).

In above scenario, the service-orientation of social information services will provide two benefits to Bella. The social information services will be accessible as traditional services along with their functional and non-functional properties. This would enable Bella to search, select and compose appropriate services for data gathering, preprocessing and information extraction based on the required features. In addition, the functional and non-functional properties will help her to predict and construct trade-offs for performance and budgetary preference such as time and cost.

4.3 Related Work

Social sensors use different types of social information services for various purposes. For instance, Twitter service is used for quick information sharing which may include short text messages, images, and videos. In comparison, YouTube service is a video sharing site that allows social sensors to upload, watch, rate, share, and add comments on videos. In addition, both social information services have different number of participating social sensors. Consequently, the data generated by YouTube and Twitter services differs in features such as quality, length, frequency, and volume. Several studies have also shown that different social information services and their features, affect the data produced by social sensors. In [Waterloo et al., 2018], researchers found that social sensor may use multiple social information services for sharing different types of data. In addition, their expressions of emotion are also diverse on social information services. Researchers analyzed the data collected from Facebook and Twitter after the 2016 Louisiana flood and Hurricane Sandy of 2012 [Kim and Hastak, 2018]. The analysis highlighted the differences between the data features (e.g., text) and the sharing patterns of social sensors. In another work [Gintova et al., 2019], investigators discovered that social sensor utilize Facebook and Twitter differently for interacting with Canadian immigration services. Moreover, they found that social sensors are likely to Tweet on Twitter instead of commenting on Facebook page, and their levels of activities are also unequal on Facebook and Twitter.

Despite of various differences, traditional analysis tools and approaches do not distinguish
between social information services and consider them at the same level. There are many online commercial and non-commercial tools developed for applications like social information service monitoring, analytics and sentiment analysis. For instance, LikeAlyzer\(^1\) is a tool designed for monitoring the performance of a Facebook pages. It provides the page analysis into different metrics such as page ranking, social sensor activities and traffic performance. Similarly, SentimentViz\(^2\) provides the analysis such as sentiment of a limited number of Tweets collected from Twitter service for given keywords. The analysis include features such as sentiment classification, 24 hour time-line, location of tweets. However, both of these tools are only developed for a particular social information service analysis and do not support the analysis of other social information services. In contrast, there are also several tools available that enable the analysis for more than one social information services. For instance, Brand24\(^3\), SocialMention\(^4\) and Hootsuite\(^5\) are online applications which can simultaneously analyze multiple social information services. However, these applications are limited in capabilities and mainly designed for customer engagement, online brand or product monitoring, scheduling and publishing of content for marketing and business. These tools provide analytical reporting dashboards which contain the information of social sensors clicks, likes, trending words or contents, general sentiment categorization, etc. Moreover, these applications treat and analyze different social information services at same level and do not allow customized analysis options.

The service-orientation is a concept which allows one to visualize computing resources as services. One example of service-orientation is the service-oriented architecture (SOA). SOA is an architectural pattern that uses services as building blocks to build new software applications [Daud and Kadir, 2014]. In order to use services in SOA based application, a software needs to be exposed as a discoverable entity with standardized interface which is accessible over the network. A service provider publishes the service information in a service registry. From the service registry, a service consumer can select a required service [Hsu et al., 2017]. Generally, a service is published with its functional and non-functional (i.e., QoS) information. Thus, a service consumer can discover services based on functional and QoS requirements. The functional aspects of a service defines the operational capabilities of a service. In comparison, the QoS is a set of non-functional attributes which provide the ability

\(^1\)http://likealyzer.com/
\(^2\)https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
\(^3\)https://brand24.com/
\(^4\)http://www.socialmention.com/
\(^5\)https://hootsuite.com/
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

to discover a service based on its non-functional properties (e.g., price) [Li et al., 2017]. The QoS attributes can be classified into different categories. In addition, these attributes allow one to quantify the performance of a service [Kaewbanjong and Intakosum, 2015].

One key ability of SOA is to define, register and dynamically compose services [Medeiros et al., 2014; Bouguettaya et al., 2017]. By leveraging the SOA paradigm, many businesses have developed new services by using existing software as services which can be used on-demand by service consumers. With the increasing number of available services, one fundamental challenge is how to efficiently and effectively manipulate services (e.g., discover, rank, compose) based on functional and QoS requirements of end users [Zhang et al., 2013]. Over the time, there are multiple frameworks and service description languages developed by researchers to describe different aspects of services and how to manipulate them [Lemos et al., 2016]. Some of the earlier and widely recognized frameworks are Universal Description, Discovery, and Integration (UDDI), Web Service Definition Language (WSDL), and Resource Description Framework (RDF). However, these frameworks only support the syntactic or keywords based service manipulation tasks such as service discovery and selection [Chen et al., 2017]. On the other hand, there are several semantic service frameworks developed which leverage the capabilities of semantic ontology for describing and manipulating services. The semantic frameworks enable service providers to define their services by using semantic annotations [Wang et al., 2015]. The semantic ontology helps to define different relationships, taxonomies and logic for services which make the manipulation process more flexible. Some prominent Semantic Web Service (SWS) frameworks are Web Ontology Language for Services (OWL-S), Unified Service Definition Language (USDL) and the Web Service Modeling Ontology (WSMO) [Klusch et al., 2016]. The semantic description frameworks have been adopted by various domains other than SOA. In [Hu et al., 2015], an agent-based multi-layer framework is developed based on context-aware semantic service (CSS) for developing context-aware applications for drivers, passengers, and pedestrians. In another work [Bermudez-Edo et al., 2016], a semantic model for sensor network to efficiently discover the sensor data is presented. In [Kulcu et al., 2016], investigators presented a survey of semantic Web and big data technologies for social information services analysis, and highlighted the potential of semantic technologies for adding capabilities to current applications.

To best of our knowledge, there is no related work available which utilizes any semantic frameworks for describing and visualizing social information services as traditional services. In this chapter, we harness the power of service-orientation to model social information services, and analyze their functional and QoS features. We formalize the service-orientated
model of social information services by using OWL-S. This would enable to conceptualize
social information services as traditional Web or cloud enabled service. As a result, our
service composition framework will be able to manipulate social information services based
on their features and compose appropriate sentiment analysis services without relying on any
classification model.

4.4 Methodology

A typical Web enabled service is generally utilized by an application, and does not have a
direct interaction with end users. In contrast, social information services have two distinct
characteristics. 1) Social information services are directly utilized by their users (i.e., social
sensors). 2) Social information services have various number of participating social sensors
which produce a large data with varying level of features. Therefore, to collectively imple-
ment the notion of service-orientation on the social information service, we define a three
step methodology. Figure 4.2 illustrates the service-orientation process of social information
services. Each methodology step is explained in the following subsection:

Abstraction

In first step, we present an abstract meta service model for visualizing the different attributes
of social information services. We then classify generic functions of social information ser-
services with regard to their data generation features. Finally, we propose a QoS model that collectively captures the non-functional features of social information services.

Conceptualization

In second step, to check the validity of our proposed abstraction model, we study the behavior of social information services by quantifying their data and QoS features. We hypothesize that the data and QoS properties of a social information service varies across diverse domains, as it is possible that the features of functionally similar services may fluctuate. For evaluation, we collected three social information services data for three diverse domains: Entertainment, Health, Politics.

Implementation

In third step, we illustrate a formal specification structure to present social information services as Web accessible services. We use Web Ontology Language for Services (OWL-S) as a baseline structure for formalization. The proposed model utilizes basic concept of semantic ontology to describe social information services and their features for automatic service manipulation.

Domains and Data Collection

To validate our proposed model, we have selected three different types of social information services for the data gathering: Facebook, Twitter, YouTube. These three services belong to different types of social information service classifications [Kaplan and Haenlein, 2010]. For the data gathering process, we have utilized open source tools developed for data harvesting from social information services. For Facebook and Twitter services, we collected data by using Facepager\textsuperscript{6}. Facepager uses ‘Graph API’ to collect Facebook posts and comments from publicly accessible Facebook pages. In contrast, Facepager uses ‘Twitter Streaming API’ to search random tweets for a given set of keywords. For YouTube service, we used an online scraper\textsuperscript{7} to collect the comments posted by social sensors under the public videos.

We selected three domains for data collection: Politics, Entertainment, and Health. The data is collected for a 7 days period between 19-March-2017 to 26-March-2017. Table 4.1 summarizes the details of the collected data for each domain\textsuperscript{8}.

\textsuperscript{6}https://github.com/strohne/Facepager
\textsuperscript{7}http://ytcomments.klostermann.ca/
\textsuperscript{8}It should be noted that despite the noticeable disparity in datasets for each domain, all of data was
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Table 4.1: Dataset Details (By Topic of Interest)

<table>
<thead>
<tr>
<th></th>
<th>Facebook (Posts)</th>
<th>Twitter (Tweets)</th>
<th>YouTube (Comments)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>108756</td>
<td>133043</td>
<td>3983</td>
</tr>
<tr>
<td>Entertainment</td>
<td>15066</td>
<td>132180</td>
<td>6961</td>
</tr>
<tr>
<td>Health</td>
<td>314</td>
<td>127984</td>
<td>66</td>
</tr>
</tbody>
</table>

- For politics domain, we selected American president ‘Donald Trump’ as topic of interest for the data collection. The collected data is based on social sensors conversations including Tweets, YouTube videos’ comments, and Facebook comments on Donald Trump’s official Facebook page.

- For entertainment domain, the newly released Hollywood movie ‘Beauty and the Beast’ was selected as a topic of interest. The data was collected from Twitter, Facebook posts from the official marketing Facebook page of the movie, and comments from fan based movie reviews YouTube videos.

- For health domain, ‘flu’ is chosen as a topic of discussion for data gathering. The data is collected from Twitter, several public Facebook pages and recent YouTube videos.

4.5 Abstraction: Social Information Services

In this section, we define the abstract model for social information services. First, we specify a high level meta-service model of social information services. Secondly, we classify social information services based on generic functions based on their capabilities of data generations. Finally, we develop a QoS model to capture the non-functional features of social information services.

4.5.1 Meta-Service Model

The meta-service model implies the highest level of abstraction for a service and its relevant properties. A social information service is defined as a tuple of five key elements: \(< ID, P, C, DT, Q >\), where

- \(ID\) is the unique service identifier.

- \(P\) is the actual social information service such as Twitter, Facebook, or YouTube.
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Table 4.2: Social Information Service Functional Classification

<table>
<thead>
<tr>
<th>Service Classification</th>
<th>Generic Functions</th>
<th>Data Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCS</td>
<td>Multimedia Content Sharing</td>
<td>Multimedia, Descriptive</td>
</tr>
<tr>
<td>SNS</td>
<td>Personalized Data Sharing</td>
<td>Multimedia, Textual, Descriptive</td>
</tr>
<tr>
<td>BLS</td>
<td>Instant Data Sharing</td>
<td>Textual, Descriptive</td>
</tr>
</tbody>
</table>

- $C$ is the classification type of a social information service.
- $DT$ defines the type of data produced by the social information service which is considered as the functional features.
- $Q$ is a set $< q_1, q_2, ..., q_n >$, where $q_i$ denotes a non-functional feature.

4.5.2 Functional Model

In previous chapter, we defined social information services based on their application usage features. In this chapter, we modify the social information service classification based on their generic functions and data produced as a result of these functions. The functional model of a social information service is comprised of two elements: Service Classification $C$, Data Type $DT$. The service classification $C$ presents the genre of a service based on generic functions. The data type $DT$ specify the data (e.g., text, video, images) generated by a social information service. Table 4.2 provides the details of functional classification of social information services.

Content Communities: Content Communities ($CCS$) such as YouTube, Instagram, and LiveLeak are services that mainly produce mainly two types of data. First type includes multimedia data in the form of videos, audios, and images. Second type comprises of descriptive data such as comments, likes, dislikes of social sensors uploaded as a response to multimedia data.

Social Networking Sites: Social Networking Sites ($SNS$) such as Facebook and LinkedIn services are Web applications that offer social sensors to create social community networks. Social sensors develop personalized profiles and develop community circles (e.g., groups, pages) for different topics of interests. Social sensors invite friends, colleagues and family members to join these community circles. In addition to multimedia and descriptive data, social sensors also produce pure textual data via text messages and status posting.

Blogging Sites: Blogging sites ($BLS$) consist of blogs and micro-blogs services. Blogs services are special types of Web-sites or online forums (e.g., Reddit, Quora) used for sharing
textual data such as questions and answers as well as articles. In contrast, micro-blogging is a new derivative of blogs which allows to share instant short textual messages. Micro-blogging services (e.g., Twitter, Tumbler) facilitate quick information sharing with friends, family and public. Other social sensor respond with same textual or descriptive data.

4.5.3 Quality of Service Model

In service-oriented architecture, a service can have multiple QoS properties to describe its non-functional features. The QoS properties provide the leverage to select and compose services based on end users’ preferences. Traditional services provide QoS features (e.g., response time, accuracy, throughput) to demonstrate their performance measures, however to determine the quality features of social information services, we propose to utilize the different characteristics of social sensor data as QoS features. These data features as QoS will allow to determine the non-functional characteristics of a social information service, and enable to compose appropriate services for its data processing and analysis. The proposed QoS features are established by investigating multiple research domains of sensor cloud computing, data quality assessment and sentiment analysis. The proposed QoS model is extensible. Currently, we focus on the following properties:

Relevancy: Relevancy property defines the extent to which the data provided by a social information service is applicable to a given domain or topic [Kevin et al., 2009]. The level of data relevancy enables to choose a suitable information extraction or noise removal service.

Richness: The social information service data is inherently diverse and rich due to the participation of multiple social sensors [Aggarwal and Abdelzaher, 2011]. Richness defines the extent of participation by unique social sensors producing data.

Volume: The volume property defines the ability to measure the quantity of data produced by social information services. It is possible that the volume of social sensor data may change after preprocessing steps. However, in composition scenarios, the volume property is a useful predictor to select scalable data preprocessing and analysis services.

Freshness: The freshness implies that the data produced by social information services is recent and does not contain any old data [Kevin et al., 2009]. Freshness provides the ability to determine data posted by social sensors with respect to temporal composition scenarios.

Spatial Information: Spatial information of social sensors is considered very useful to visualize their data based on geo-locations [Gao et al., 2011]. However, not every social information service provides social sensors to expose their locations via mechanism like geo-
tagging. The spatial information property specifies the ability of social information service to expose social sensors data with their locations. As a result, in the absence of spatial information, a location extraction service can be composed to extract the social sensor locations during composition scenarios.

**Text Type:** There are two types of text writing styles in online data sharing: formal, informal [Thelwall et al., 2010]. In informal style, the written text is influenced by the Internet language which includes different abbreviations, slang, and emoticons. In contrast, by using formal style, social sensors utilize proper language skills, and avoid Internet language. The text type property classifies the collected textual data into two types of writing styles. The text type can be used for appropriate sentiment analysis service selection.

**Lexical Diversity** Social information services often restrict social sensors for writing text by limiting the number of characters to be used. The lexical diversity [Baldwin et al., 2013] determines the lexical heterogeneity of textual data in terms of the number of characters, words and sentences. Lexical diversity can also be used for appropriate sentiment analysis service selection.

**Meta-Data Properties:** Social information services provide social sensors a range of options for responding to a content shared by other social sensors. These options present the different types of descriptive data (i.e., meta-data) related to the shared contents [Zeng et al., 2010]. The meta-data properties present the actions committed by social sensors such as number of likes, shares, and downloads on shared data. These properties can supplement the social sensor data.

### 4.6 Conceptualization: Functional and Non-Functional Analysis

In this section, we analyze and quantify the functional and QoS properties of social information services. We investigate our dataset with respect to three different topics, and demonstrate the analysis results.

#### 4.6.1 Functional Properties Analysis

For functional analysis, we breakdown our dataset into 7 days interval for each topic per social information service. For each day, the number of data contents and respective responses are recorded. The analysis shows that both Facebook and YouTube services have notable variations for data generation for different topics. Social sensors are more active for entertainment and politics domains. In comparison, for health domain social sensors are sharing
limited amount of data. On the contrary, Twitter service remains almost consistent with the flow of the data generation by social sensors in three domains. The main reason for its consistency is the mechanism provided for gathering tweets by using ‘Streaming API’ which allows the tweet collection from Twitter service as a single entity. Unlike, the other two social information services which do not offer a standardized mechanisms to collect the data by treating them as a whole, rather the data can be only collected from individual Facebook pages and YouTube videos. Therefore, it validates our hypothesis that different types of social information services produce data with varying types, frequencies and volumes based on diverse domains. Table 4.3 presents the analysis results of social sensors shared data and related responses.

4.6.2 Non-Functional Properties Analysis

For each of the proposed non-functional properties, we applied various techniques to measure the QoS features. In our analysis, we analyze the QoS features in the defined order.

**Relevancy:** For removing the irrelevant data and noise from our dataset, we used the inclusion and exclusion technique [Injadat et al., 2016]. We have used following three data
Figure 4.3: Irrelevant Data In Three Social Information Services by Topic

filters to retain the relevant textual data:

- **Language Filter**: is used to exclude the data based on a natural language. We have excluded all non-English data.

- **Search terms Inclusion Filter**: is used to include the data based on keywords. For instance, we used “Trump” and “US President” to include the relevant data items for politics. For the sake of standardization, we used same search keywords for each domain across the dataset for the data inclusion.

- **Stop Words/Phrases Exclusion Filter**: is used to exclude a data item in dataset which is irrelevant. For instance, a data item containing “Ivanka Trump” may contains the keyword “Trump”. However, it is not relevant to the context of ‘President Trump’. Hence, “Ivanka” is utilized as a stop word to exclude such data items. Although the selection of search and stop words is exclusively dependent on the domain and extent to which the data relevancy is desired.

The above filters can be applied in different sequences and combinations. For Twitter service, we applied the above three filters. We found that the Twitter service data contains a large amount of irrelevant data. One main reason is that the data provided by Twitter API is randomly sampled based on keyword matching rather than the contextual matching. For example, the health data contained a large number of tweets which include the term “Beiber Fever” by the fans of singer Justin Beiber. Consequently, much of the data is discarded by using stop word ‘Beiber’. Moreover, the maximum bulk of the tweets are discarded due to written in non-English languages.
For Facebook and YouTube services, we used the 1st and 3rd filters. Unlike Twitter service, the data on Facebook and YouTube services is directly posted by social sensors as response under the content (e.g., video). For example, the comment “Awesome Movie:)” posted on the ‘Beauty and the Beast’ Facebook page has no direct search term, still it is relevant to the context. Therefore, we did not apply the search filter. We find out that a larger portion of data on both Facebook and YouTube services is written in English language. Consequently, overall relevancy of Facebook and YouTube services is greater than Twitter service. Figure 4.3 presents the results of irrelevant data percentages of each service by domains.

**Richness**: The different functionality of social information services attract various ranges of social sensors for online participation. For each social information service, the participation of unique social sensors may vary with respect to a particular domain. Figure 4.4 presents the number of unique social sensors by topic for each social information service. For Twitter service, the number of unique social sensors are extracted based on the social information service generated identifiers. Due to the unavailability of unique identifiers in the datasets of Facebook and YouTube services, we used the names of social sensors as an alternative of unique identifiers. The duplicate entries of social sensors are not accounted. In our analysis, we observed that a large number of social sensors are active on official Facebook pages. Moreover, Facebook service data has a higher number of participants for politics and entertainment domains in comparison to Twitter and YouTube services. On the other hand, Twitter service has more participant social sensors for health data than Facebook and YouTube combined. Unexpectedly, YouTube service has the lowest social sensors, in terms
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Figure 4.5: Data Time-line in Social Information Services by Topic

of comments sharing for all three domains.

**Volume and Freshness:** Social sensors share their data in two modes: Stream and Non-stream. In stream mode, social sensors continuously share their data in limited time related to any recent event. In non-stream mode, social sensors passively share data over the time in response to on-going events. We define the freshness of our dataset by time slicing with an interval of 7 days. However, the time slicing of data can be done based on different time parameters (e.g., hours). Due to the nature of Twitter API, the Twitter service data shows consistent streaming behavior for data generation within defined time parameters for three domains. In comparison, the data on both Facebook and YouTube services is generated based on a sudden event in stream mode, or on a simple shared content in non-stream mode. Both services have shown high variations for three domains with respect to time parameters. We combine the visualization of the volume property of filtered data with its freshness. Figure 4.5 shows the shared data time-line for 7 days period based on social information services by each domain.

**Spatial Information:** The spatial information reveals the origin of the data by using social sensor locations. Some social information services (e.g., Twitter) provide the facility to social sensors for exposing data with geo-location information (e.g., geo-coordinates). However, several services (e.g., Facebook and YouTube) do not provide such a facility. For such social information services, the location information can be extracted through text parsing by checking the mentions of the location names (e.g., cities, towns). The location parsing is a probabilistic approach which may not give exact location of a social sensor. However, it can be used as a minimal option for location information extraction. Currently, we assume the mentions of locations in text with conjunction of terms ‘in’ and ‘from’ as social sensor
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Table 4.4: Signal to Noise Ratio (SNR) of Spatial Information

<table>
<thead>
<tr>
<th></th>
<th>Facebook</th>
<th>YouTube</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>0.387</td>
<td>0.401</td>
<td>0.011</td>
</tr>
<tr>
<td>Entertainment</td>
<td>0.069</td>
<td>0.16</td>
<td>0.046</td>
</tr>
<tr>
<td>Health</td>
<td>0.048</td>
<td>0.087</td>
<td>0.054</td>
</tr>
</tbody>
</table>

SNR = Signal/Noise
Signal = Geo-Tagged Data Items, Noise = Total-Signal

Table 4.5: Formal To Informal Data Ratio

<table>
<thead>
<tr>
<th></th>
<th>Facebook</th>
<th>YouTube</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>1 : 0.08</td>
<td>1 : 0.07</td>
<td>1 : 6.45</td>
</tr>
<tr>
<td>Entertainment</td>
<td>1 : 0.13</td>
<td>1 : 0.16</td>
<td>1 : 2.04</td>
</tr>
<tr>
<td>Health</td>
<td>1 : 0.08</td>
<td>1 : 0.32</td>
<td>1 : 1.76</td>
</tr>
</tbody>
</table>

locations. We used the Stanford NER (Named Entity Recognizer) for location extraction. The Precision, Recall and F-Measure of Stanford NER with baseline training dataset is 0.699, 0.682 and 0.691, respectively [Lingad et al., 2013]. Although Stanford NER is a useful tool for location information extraction; sophisticated semantic or text analysis techniques are required to determine the exact location names of social sensors. For politics domain, all social information services have the highest number of geo-location information. Meanwhile, all three services have the lowest geo-location information for health domain. We used Signal to Noise Ratio (SNR) as an evaluation metric to present the ratio of geo-location in our dataset. Table 4.4 presents the spatial information SNR of Twitter service based on the data items with geo-coordinates, and the SNR for Facebook and YouTube services based on NER location extraction.

Text Type: We used a binary classifier to classify the data into two categories: Formal, Informal. Our classifier utilizes a set of words obtained from Internet slang dictionary\(^9\). It contains trending acronyms, slang and abbreviations such as lol, imho, and omg, which are used by social sensors on different types of social information services. Based on the classification results, we find that Twitter service has the highest ratio for informal data for all of the domains. Social sensors using the Twitter service tend to use more slang, and abbreviations. In contrast, social sensors of both Facebook and YouTube services have higher tendency to post their data in formal style. Table 4.5 shows the classification ratios of text types for three social information services per domain.

Lexical Diversity: We analyze the lexical diversity of data based on three measures: Char-

\(^9\)https://www.internetslang.com/
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Table 4.6: Lexical Analysis of Social Information Services by Topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>Measure</th>
<th>Facebook</th>
<th></th>
<th>YouTube</th>
<th></th>
<th>Twitter</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Min</td>
<td>Max</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Politics</td>
<td>Characters</td>
<td>1</td>
<td>8935</td>
<td>1</td>
<td>3975</td>
<td>5</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>Words</td>
<td>1</td>
<td>1446</td>
<td>1</td>
<td>723</td>
<td>5</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Sentences</td>
<td>0</td>
<td>86</td>
<td>0</td>
<td>53</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Entertainment</td>
<td>Characters</td>
<td>1</td>
<td>7252</td>
<td>1</td>
<td>3478</td>
<td>11</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>Words</td>
<td>1</td>
<td>1463</td>
<td>1</td>
<td>651</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Sentences</td>
<td>0</td>
<td>85</td>
<td>0</td>
<td>47</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Health</td>
<td>Characters</td>
<td>5</td>
<td>916</td>
<td>25</td>
<td>573</td>
<td>5</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>Words</td>
<td>1</td>
<td>173</td>
<td>4</td>
<td>112</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Sentences</td>
<td>0</td>
<td>9</td>
<td>1.2</td>
<td>0</td>
<td>17</td>
<td>2.6</td>
</tr>
</tbody>
</table>

The measures are computed by using three functions: Minimum, Maximum, Average. Before extracting the measures from our dataset, we first truncate all data items from online markup such as hashtags, user mentions, emoticons and special characters. Secondly, we also accommodate out-of-vocabulary (OOV) words [Aggarwal and Abdelzaher, 2011] in our analysis. We do not discard misspelled words, slang, and abbreviations. In our analysis, we found that due to 140 characters limit for Twitter service, the lexical properties are almost consistent and there is not significant impact of domains on these lexical features. Both Facebook and YouTube services have largely similar results for minimum function, where social sensors may only post a simple character (e.g., ‘D’, ‘P’) or word (e.g., cool, awesome) as a response to the content. However, both services show variations for maximum and average functions due to the larger character limit. Social sensors may use a large number of characters while posting comments on Facebook service as well as on YouTube service. Table 4.6 presents the summary of lexical diversity analysis of social information services by three domains.

**Meta-Data Properties:** Social information services have multiple types of meta-data properties. To simplify, we classify meta-data properties into two groups: Reaction, Propagation Index (P-Index). Reaction depicts the complimentary actions of social sensors which show their emotions and sentiments in terms of likes, and dislikes in response to a content. Propagation Index describes the dissemination of a content like number of shares, views, and downloads by social sensors. We only analyzed a subset of our dataset to extract the meta-data properties. It is noticed that the videos on YouTube service have less reactions and comments, however, social sensors have high P-Index for the videos. Moreover, it is also observed that social sensors are more likely to post their reactions instead of sharing their comments. Meanwhile, Twitter service has one reaction property (i.e., Like), shows differ-
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Table 4.7: Meta-Data Analysis of Social Information Services

<table>
<thead>
<tr>
<th>Domain</th>
<th>Facebook # of Reactions</th>
<th>P-Index</th>
<th>YouTube # of Reactions</th>
<th>P-Index</th>
<th>Twitter # of Reactions</th>
<th>P-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>L\149,000 V\13,200 H\3048, W\436 S\262, A\534</td>
<td>11,442</td>
<td>L\3208 D\158</td>
<td>70,014</td>
<td>L\21,753</td>
<td>10,171</td>
</tr>
<tr>
<td>Entertainment</td>
<td>L\106,800 V\26,400 H\3313, W\383 S\19, A\30</td>
<td>9939</td>
<td>L\2179 D\45</td>
<td>170,159</td>
<td>L\18,200</td>
<td>7876</td>
</tr>
<tr>
<td>Health</td>
<td>L\144, V\3 H\0, W\5 S\4, A\1</td>
<td>162</td>
<td>L\155 D\2</td>
<td>2272</td>
<td>L\8291</td>
<td>4436</td>
</tr>
</tbody>
</table>

Notations*: L=Likes, D=Dislikes, V=Love, H=Haha, W=Wow, S=Sad, A=Angry

A large amount of reactions for three domains, and a limited number of tweet sharing. On the contrary, since Facebook service provides a range of option for social sensors to register their reactions, it has more rich and divers reaction information. Although social sensors show a large number of mixed reactions, its P-Index is comparatively limited. Table 4.7 summarizes the results of meta-data analysis for three social information services by each domain.

### 4.7 Implementation: Social Information Service

In this section, we demonstrate the applicability of a Web accessible service model for social information services. For service composition scenarios such as Bella’s, she has to find and compose appropriate services based on end user requirements. Generally, a service provider publishes the information of a Web based service that allows potential service consumers to discover the service. We present a formal specification structure to define a social information service as Web accessible service by using OWL-S. The OWL-S is a Web markup language which enables the development of semantic ontologies to present services and their contents on the Web. The semantic ontology mechanism provides the flexibility for end users and various agent based applications to discover, invoke, compose, and monitor services in a automated environment. The main advantage of using OWL-S ontology model is that it encapsulate the existing and well established building blocks (e.g., WSDL, XML) in the Web enabled service technologies. Moreover, it is being evolved under the standardized W3C’s Web Services Description Working Group [Martin et al., 2007]. The OWL-S ontology
structure (see Figure 4.6) has three main components: Service Profile, Process Model, Service Grounding. The service profile defines the core functional capacities and non-functional properties of a service, and elaborates that what the service offers. The service model presents the detailed information of the internal working process of a service and demonstrates that how the service achieves its functionality. The service grounding provides details of technical specification such as service end points, communication protocols, transport mechanisms, and data formats required to invoke the service.

In the remainder section, we present the social information services by using the OWL-S ontology models and their topologies. For elaborating three service ontology models, we use Twitter service as an example of social information service.

4.7.1 Service Profile

The service profile provides the high-level descriptive information of a service in textual format. The service profile is created as an end user readable description of the service. The service profile includes the number of functions which the service can perform, and non-functional properties declared by the service. The functional and non-functional properties are used to define semantic descriptions by service providers to publish the service information. This information helps to discover a relevant service by the end users. In OWL-S, a service profile is defined as a abstract class which represents the service descriptions. The ser-
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Figure 4.7: Service Profile Class

Service profile class can be divided into three types of elements: Service Information, Functional Description, Service Attributes. Each service element is further divided into sub-elements. Figure 4.7 presents the overview of a service profile class. The details of service profile elements and their structure is provided as follows:

- Service Information: provides a human readable information of the service including service name (e.g., Twitter service), service text description such as its introduction, nature of usage, type, etc., and the contact information of service provider.

- Functional Description: contains the information of all functions that are offered by the service, and it specifies the conditions for their successful execution. For instance, Twitter service provides function for collecting tweets and how this function can be executed. The functional description includes the required inputs parameters to invoke the service, and the outputs produced by the service. It also provides the preconditions which are required to be met before the service is invoked. For example, for collecting tweets, the user credentials must be valid. Finally, it defines the potential effects that may occur as a result after the service execution (e.g., function success).
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

- Service Attributes: are further divided into two types of attributes: Service Category, Service Parameters. The service category attributes identify the generic classification of a service. The service parameters has the list of QoS properties offered by the service. For instance, Twitter service can provide social sensor data based on freshness and locations information.

4.7.2 Service Model

The service model provides the information to an end user of how to interact with the service. The service model can be viewed as a process (i.e., service process). The service process shows the workflow of service function(s) which provides the step by step specification of multiple ways for an end user to interact with the service functionality. It is possible that a service may have an atomic process or a complex process. An atomic process depicts that a service receives one input message, and subsequently it returns one output message. On the other hand, a composite process maintains state and moves through the process based on different end user’s messages. For example, a typical Twitter service search process is initiated by validating the end user’s credentials. If the credentials are not verified by the service, the search process is aborted. Otherwise, the search process advances to the next state where the input (i.e., search keywords) provided by the end user are validated (e.g., input length). Similarly, if the input is rejected, the search process is terminated. On the other hand, if the input is successfully validated, the search process returns the tweets matched with input, and finally the process is successfully completed. Figure 4.8 shows the composite process model of Twitter service to search the relevant tweets for a given input.

A service process has two main tasks: Information Production, Transition. The information production task depends on two parameters: Inputs, Outputs. The transition task

Figure 4.8: Composite Service Process Model: Search Tweets
is defined by two parameters: Preconditions, Effects. Collectively, the parameters: Inputs, Outputs, Preconditions, Effects, are known as IOPEs. The IOPEs are explained as follows:

- **Inputs**: establish the parameters required to invoke a service process. It is possible that a service requires multiple number of inputs or none. For instance, the *Search Tweets* process requires user credentials and search keywords.

- **Preconditions**: are the prerequisite conditions which are expected to be true prior the successful execution of service process (e.g., validation of inputs). There can be several preconditions linked with a single process.

- **Outputs**: are the information produced after the execution of the service process. A service may produce multiple number of output messages for service user. For instance, the *Search Tweets* process may provide a number of matching tweets.

- **Effects**: present the different outcomes of the service process depending on the preconditions required by the process (e.g., failure, success messages).

Although a service defined by using OWL-S does not necessarily require to model a service process, however, it must have a service profile and service grounding in order to be accessed over the Web.

### 4.7.3 Service Grounding

The service grounding supplies the information of how to access the service. The service grounding is mainly used to define the technical details such as communication protocol (e.g.,
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

Figure 4.10: Twitter Service Grounding For Twitter Streaming API
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

HTTPS), data transportation protocols (e.g., SOAP), message format structures, and service end points. The service grounding is regarded as the concrete specification of the service, whereas the service profile and service model are considered as abstract representation of the service. Therefore, the service grounding is constituted as the concrete binding of the service with respect to its abstract representation. The concrete specification of the service grounding can be presented by different service description languages and technologies.

It is possible that there are multiple concrete service representations available for a service grounding. For example, Twitter as a service can be accessed through Twitter service provided APIs (e.g., Twitter Search API, Twitter Streaming API). On the other hand, there are many third party customized services developed by using Twitter service provided APIs. Third party services such as TweetDeck\textsuperscript{10} and Tweetinvi\textsuperscript{11} can also be presented and accessed as concrete service bindings. Moreover, concrete service bindings can be a domain specific. For instance, a customized service may only provide Twitter service data for politics domain. Figure 4.9 presents the deployment model of Twitter service with multiple groundings. Figure 4.10 depicts the concrete service groundings representation for Twitter Streaming API by using XML structure.

4.8 Chapter Summary

Social Information services provide different abilities to produce data with diverse features. These features are impacted by different domains or topic of interests such as politics, entertainment, etc. However, current social information service analysis tools consider the social information services and their generated data as homogeneous entity, and do not differentiate between the changing feature of social information services. In this chapter, we implemented a novel strategy to capture and analyze the diversity of social information services by harnessing the power of service-orientation. We presented a meta-service model to visualize social information services as Web based services. We classified social information services based on their functional features. We proposed a novel QoS model to capture the non-functional properties of social information services. To validate the applicability of our proposed approach, we collected the data from three different types of social information services: Facebook, YouTube, Twitter, for three dissimilar domains: Politics, Entertainment, Health. We conducted a series of experiments to quantify the functional and QoS properties.

\textsuperscript{10}https://tweetdeck.twitter.com/
\textsuperscript{11}https://tweetinvi.codeplex.com/

86
CHAPTER 4. SOCIAL INFORMATION SERVICES: THE SERVICE-ORIENTATION OF SOCIAL MEDIA

The analysis results strengthened our hypothesis, and demonstrated the varying levels of social information services features and impact of different domains on these features. Finally, we formally presented the social information services as Web based services by using the Ontology Web Language for Service (OWL-S). As a proof of concept, we used OWLS-S ontology models to specify different aspects of Web enabled social information services.

In next chapter, we integrate the proposed QoS model in our existing service composition framework to dynamically compose sentiment analysis services for large scale social information service data.
Chapter 5

QoS Driven Service Composition
For Big Social Data

Social information services hold a prominent position as a member of the big data family. Social information services provide free and robust infrastructure, where social sensors share data at an unprecedented speed without being concerned about storage and processing. Consequently, the data produced by social information services is growing exponentially, and this data has various dynamic features. Current techniques for analyzing large amounts of social information services data do not consider their changing features, and treat social information services as a uniform data source with similar features. In Chapter 4, we demonstrated that social information services have diverse features. Also, we developed a QoS model that is capable of capturing various features of social information services.

In Chapter 3, we developed a composition approach that used social information services classification model to find the appropriate sentiment analysis services. However, the classification model may not be an ideal solution for composing sentiment analysis services for unpredictable features of large scale social information service data. In this chapter, we integrate the QoS model into our existing sentiment analysis service composition framework. The QoS model replaces the existing social information service classification model as a dynamic mechanism for service composition. We develop a new service composition technique based on graph-planning that dynamically assesses the QoS features of social information service data, and composes appropriate services required for sentiment analysis. We present social information service based sentiment analysis system as a motivating scenario. We conduct a set of experiments on real-world datasets and demonstrate the efficiency of our approach.
CHAPTER 5. QOS DRIVEN SERVICE COMPOSITION FOR BIG SOCIAL DATA

5.1 Introduction

The term of big data is used to describe the large amounts of complex datasets which are beyond the computational abilities of traditional tools and techniques to collect, manage, process and analyze within reasonable time-frames [Becker et al., 2015]. Although big data is not a novel concept, there is no clear definition of ‘Big Data’. Instead, it is defined on the basis of its three characteristics or ‘3Vs’: Volume, Variety, and Velocity [Zaslavsky et al., 2013]. Social information services have retained a unique status as a part of big data family. Due to cheap Internet access and free accessibility by social sensors, social information services have become a leading force in online data generation. Moreover, social information services have emerged as free sources of big data which is desirable in different applications.

The large quantity of data which is specifically generated from social information services is defined as ‘Big Social Data’ with similar 3Vs properties described as follows:

- Social Information Service Volume: According to statista.com, by the end of year 2021, there will be 3.02 billion social sensors sharing their data online. Therefore, the increasing number of social sensors results in exponential growth of social information service data size.

- Social Information Service Variety: There are various types of social information services such as blogs and micro-blogs, social networking sites, and content communities [Kaplan and Haenlein, 2010]. The data produced by these services is unstructured and available in several formats such as audio, video, text, etc.

- Social Information Service Velocity: Social information services generate data at an unprecedented speed. For instance, in 2014 only, Twitter service reported 500 million tweets published by social sensors per day [Haustein et al., 2016]. According to Brandwatch1, as of June 2016, 95 million posts were shared per day on Instagram service. Moreover, 350 million photos are uploaded on Facebook service per day.

The 3Vs characteristics of social information services make sentiment analysis more challenging task. Despite the presence of ‘Variety’ in social information services, existing analysis techniques and tools assume the entire cyberspace of social information services as a single entity which has similar data features. In addition, these tools provide different capabilities for performing analysis and visualizing the results. Consequently, multiple tools generate

1https://www.brandwatch.com
different views of social information service data, and further complicate the aggregation of analysis outcomes. Meanwhile, the data-oriented approaches are slow and require time-consuming tasks (e.g., dataset labeling, algorithm training) to deal with the changing features of social information services data. In Chapter 3, we developed a service composition framework that uses the generic features to differentiate between social information services and composes services required for spatio-temporal sentiment analysis. However, generic features may not be applicable for composing sentiment services for big social data. Therefore, it is paramount to consider the dynamically changing features of big social data for service composition. In Chapter 4, we developed a novel QoS model that captures the dynamic features of social information services by modelling them as typical Web based services.

In this chapter, we incorporate the new QoS model in our existing service composition framework. The QoS model would enable the dynamic composition of appropriate services based on the changing features of big social data generated by social information services. The major contributions of this chapter are as follows:

- A dynamic mechanism to capture the changing features of big social data produced by social information services.
- A QoS driven service composition mechanism based on graph-planning which composes required sentiment analysis services for multiple social information services in parallel before aggregating the results.

The rest of the chapter is organized as follows. Section 5.2 presents the motivating scenario. Section 5.3 highlights the related research efforts. Section 5.4 presents the details of the solution approach. Section 5.5 provides the details of our experiments by using the real-world dataset and evaluation results. Section 5.6 concludes the chapter.

5.2 Motivating Scenario

Let us assume that ‘Jay’ is a social information service analyst who is working for a political party ‘My Nation First’. The party is heading towards the upcoming presidential election. In order to define the vision and political agenda for the election campaign, the party is interested in obtaining the general public opinion and sentiment about presidential candidates including the current president. The party wants to know how well their presidential candidates are viewed by the potential voters based on their various political and policy
standings. Furthermore, the party is keen to visualize the overall public opinion based on voting constituencies within the country for each candidate.

In this scenario, Jay is given the task for developing a sentiment analysis system for presenting the public opinion from popular social information services. Jay develops the sentiment analysis system by using a typical sentiment analysis approach in five steps (see Figure 5.1). Jay develops the system by composing several processes as follows: He first selects a set of social information services as data sources for collecting the public opinion. In particular, Jay chooses social information services which can provide opinions along with geo-locations. For each selected social information service, he defines the data preprocessing flows to remove noise, irrelevant and non-geo-tagged data by using several tools and services. Afterwards, Jay takes subsets from each dataset collected from selected social information services. Next, Jay trains one single machine learning algorithm for data analysis (e.g., sentiment analysis) with all subsets. Otherwise, he may select a number of algorithms for individual social information services, and train and validate them until the desirable accuracy is achieved. Finally, after the information extraction process, different outputs of the algorithms are collected, standardized in cohesive formats and presented by using various visualization objects (e.g., charts, maps). The above approach has following limitations:

- One uniform mechanism for data preprocessing may not be applicable to entire set of social information services selected by Jay. He needs to devise a series of separate data
preprocessing steps for each social information service.

- Similar to data preprocessing, one algorithm may not be suitable for all types of social information services. Regardless of the number of algorithms chosen; over time training and validation of algorithms are needed for retaining the desirable level of accuracy for changing features which is time consuming and requires manual effort.

In this scenario, the proposed service composition framework will provide two benefits over the above traditional approach to Jay. First, the composition framework dynamically extracts and assesses the features of an individual social information service. This would allow to understand the required preprocessing and analysis requirements of a social information service. Secondly, for each social information service, it composes the appropriate services (e.g., preprocessing, data analysis) based on relevant features. As a result, Jay does not need to perform laborious and time consuming tasks.

5.3 Related Work

Big data is collectively used to denote data intensive technologies which are emerging as a new technology trend in research and industry [Demchenko et al., 2013; Sivarajah et al., 2017; Oussous et al., 2018]. Big social data is an integral part of big data [Blazquez and Domenech, 2018]. Big social data inherits its complex processing and analysis challenges from big data. In addition, big social data some has its own unique challenges which are connected to interdisciplinary areas of research such as data mining, graph mining, machine learning, statistics, natural language processing, and information retrieval. Cloud services provide an ideal platform which has the infrastructure and computational power to cope with the various processes involved in large amounts of data collection, storing, processing and visualization [Demchenko et al., 2014; Yang et al., 2017].

Regardless of impressive processing capabilities of cloud computing, there is limited work available in the context of big social data based sentiment analysis. For instance, in [Zheng et al., 2013] researchers describe an abstract paradigm for Big Data Analytics Software-as-a-Service (BDASaaS) which assumes the software as a service (SaaS) model to extract real-time information from structured and unstructured data. The BDASaaS paradigm gives a basic overview of how the large streams of data can be processed. However, it does not provide any details regarding extracting information such as sentiment and opinion. In another effort [Taher et al., 2017], a platform: Big Data Lab as a Service (BDLaaS) is designed which offers
CHAPTER 5. QOS DRIVEN SERVICE COMPOSITION FOR BIG SOCIAL DATA

to host big data analysis applications. Although the applications (e.g., sentiment analysis) can be deployed on BDLaaS infrastructure, it only supports the selection and composition of services in the form of Platform as a Service (PaaS). There are also several research efforts focusing on big data analysis by utilizing cloud based services. For instance, researchers proposed a high level big data analytics service-oriented architecture (BASOA) based on traditional SOA paradigm [Sun et al., 2018]. However, the proposed architecture does not provide any implementation details regarding how the big data is analyzed. In [Puthal et al., 2016], a cloud powered framework is developed that supports the event detection in emergency situations from data stream originating from multiple sources. The framework is designed for a secure and efficient data gathering, data integration and alert dissemination. In [Ardagna et al., 2018], a model based big data analytics as a service is designed which uses the OWL-S defined ontology for selecting and composing big data analysis services. However, mainly these approaches do not differentiate between big data and big social data, and conceptualize them equally. In addition, these approaches do not provide the means to identify and compose services specific to sentiment analysis applications.

On the other hand, there are various commercially developed tools such as Hootsuite\(^2\) and Klear\(^3\) which are used for analyzing social information services. However, these tools are mainly developed for concurrent engagement with online users and for reporting, scheduling and publishing of contents such as promotional videos, images, and interactive messages. Their application for sentiment analysis or opinion mining is restricted to limited volume of data. Meanwhile, there are a number of Web services and APIs specifically developed for sentiment and opinion extraction from text and hosted by cloud service providers [Gitto and Mancuso, 2017; Gupta and Kumar, 2017; Valdivia et al., 2017]. However, these Web services and APIs are stand alone solutions which can be utilized by developers based on their application needs. One limitation of using online commercial Web services is that they provide outputs in non-standardized formats. As a result, the usage of multiple services in a single application raises the integration challenges for converting the different outputs into one format.

Graph-Planning is an Artificial Intelligence (AI) based planning approach which takes input in terms of states and a set of constraints to determine the best strategy for reaching the desired goal state [Blum and Furst, 1997]. The AI based approaches have been successfully applied to service composition problems [Rodrguez et al., 2016]. By using the AI

---

\(^2\)https://hootsuite.com/
\(^3\)https://klear.com/
based planning, a service composition problem is visualized as a planning problem where the goal is to find suitable services based on end users’ constraints. However, there is limited research work done for enabling service driven big data analytics using AI planning. For instance, an an ontology-based workflow generation method is proposed for service composition planning for data mining applications. Ontology and rules are designed to orchestrate the assembling of data analytics processes by using the properties of a given dataset [Kumara et al., 2015]. Researchers developed two service composition algorithms which use semantic ontology for creating service composition graphs for supporting big data applications [Huang et al., 2017]. In another work [Siriweera et al., 2017], researchers developed a graph-planning approach which composes various big data analysis processes to formulate the automated workflow based on dynamic constraints. However, the proposed approaches are focused on generalization of big data analysis instead of big social data analysis. The approaches utilize the assembling of different processes for analyzing large volumes of data rather than composing software services. In our proposed approach, we employ graph planning that uses QoS features (e.g., text type) of social information services as constraints for composing services to extract sentiment from big social data.
5.4 Extended Sentiment Analysis as a Service Framework

In this section, we present the extension of our Sentiment Analysis as a Service (SAaaS) framework. In Chapter 3, we defined multiple service layers for sentiment analysis where service composition layer selects services from different layer based on social information services classification model. In the extended framework, we add two components in the service composition framework. Figure 5.2 shows the extended architecture of the framework. The first component is the Data Model which defines the data processing pipelines. A data pipeline processes only one social information service at a time for sentiment analysis by various processing and analysis services. The data model can simultaneously create multiple parallel data-pipelines for a set of social information services. Later, the results of each data pipeline are aggregated. In previous framework, the service composition layer composed services based on static social information service classification features. In the extended framework, the data model collects the data of social information services in a temporary database for features assessment. The second component extension is the inclusion of two modules: Dynamic Feature Assessment and Social Information Service Quality Model, in Service Composition Layer. The dynamic feature assessment component extracts and analyzes the QoS features of the big social data stored in temporary database using quality model, and the composition layer aggregates appropriate services for each data pipeline.

In the following two sub-sections, we present the details of the data model and the service composition approach used in service composition layer.

5.4.1 Data Model

The data model consists of a data pipeline structure for big social data analysis. A data pipeline consists of four existing service layers: Data Collection, Data Pre-processor (i.e., noise removal), Location Extractor, and Data Analyzer. In this extension, we add a new service layer (i.e., Integrator) which combines the outcome of all data pipelines. Each data pipeline enables the concurrent processing of multiple social information services based on their QoS features. In a typical scenario, a social information service is processed in following sequence. First, the number of social information services are selected based on the end user’s request. The data is collected for each social information service and stored in a temporary database. Next, the service composition layer analyzes the QoS features from the stored data, and it selects appropriate services for each data pipeline. Finally, the integrator aggregates the output of data pipelines, and delivers the analysis results to end users. In Chapter 3,
we presented the definitions of service layers and their component services. In the following section, we revise the definitions of service layers, and respective component services.

- **Definition 1:** The extended version of Sentiment Analysis as a Service (eSAS) is a composite service which integrates the component services to analyze big social data for sentiment analysis and provide results to the service users. eSAS is a tuple of 
  \(< ID, D_o, SK, SN_i, F, D_L - S_L, S - T, Q_i >, \) where

  - **ID** is the unique service identification number.
  - **D_o** is the topic of interest (e.g., politics, health) for which analysis is required.
  - **SK** provides the set of keywords or search terms for data gathering.
  - **SN_i** is the desired set of social information services to be included for analysis.
  - **F** is the set of various functions offered by the service for sentiment aggregation, comparison and presentation.
  - **D_L - S_L** defines the optional requirements for including minimum amount of data required for the analysis per **SN_i**, where **D_L** establishes the limit for minimum number of data items, and **S_L** for minimum number of unique social sensors.
  - **S - T** defines the spatio-temporal requirements to gather the data, where **S** provides the requirements of targeted geographical area. **T** represents the temporal bounds for which the analysis is required.
  - **Q_i** is a tuple of \(< q_1, q_2, ..., q_n >\) non-functional features where \(q_i\) denotes the unique QoS property such as price, response time, throughput, etc.

- **Definition 2:** The data collection service layer contains multiple data gathering services for different social information services. These services represent the concrete realization of social information services. The data collection services collect the relevant data based on user provided keyword(s) and save the data into a temporary database. A data collection service **DCS** is a tuple of \(< cid, sis, ts-te, Loc, qi >, \) where

  - **cid** is the unique identifier number of a data collection service.
  - **sis** is the targeted social information service for retrieving data based on **D_L - S_L**.
  - **ts-te** determines temporal bounds **ts/te** (start/end) time for data collection which are driven from composite service parameter **T**.
– *Loc* present the list of locations for which the data is to be extracted. The locations are defined by the user of the composite service *eSAS* by using parameter *S*.

– *qi* is a list of QoS properties (e.g., response time) of data collection service.

**Definition 3:** The Data pre-processor service layer comprised of various filters for removing different types of noise and irrelevant data. We currently utilize three filters in the following order:

– Language Filter: filters data based on a required natural language (e.g., English).

– Extra Data Filter: removes the unnecessary data such embedded URLs, repeated item and data without temporal or time stamps.

– Relevant Data Filter: The degree of relevant information in a dataset can vary for each social information service. This filter employs a set of candidate retrieval techniques (e.g., relevance, ranking) [Grossman and Frieder, 2012] which can be applied based on the quality of the data. We use a probabilistic method to assess the relevance quality of each datasets for selecting appropriate filtering technique.

A data preprocessing filter is represented as a service *PRS* which is defined as a tuple of \(< pid, di, tn, qi >\), where

– *pid* is the unique identifier number of a noise filtering service.

– *di* is a set of dataset with data items \(< d_1, d_2, ...., d_n >\) where each *di* represents a data item such as tweet, comment or post.

– *tn* defines the filter type and its functions.

– *qi* defines the non-functional features (e.g., accuracy) of noise removal filter.

**Definition 4:** The location extractor service layer extracts the location of social information services. To determine the locations of a social sensor, we use two different strategies. 1) If a service allows the geo-tagging facility, simply the geo-coordinates (i.e., longitude and latitude) in the data are used as social sensor location. Alternatively, social sensors profile information can be used as likely measure for determining their location. 2) The location information is extracted by checking the mentions of location names (e.g., cities) in the text and geo-coding technique is used to assign the geo-coordinates. The data items without location information in a dataset are eliminated. The location extractor service *LES* is defined as a tuple of \(< lid, ds, Loc, qi >\), where
\textit{Definition 5:} The data analyzer service layer contains analysis services for extracting subjective information (e.g., sentiment, emotions, opinions) from the collected data. A data analyzer service $DAS$ is presented as a tuple of $<\text{aid}, \text{op}, \text{ln}, \text{qi}>$, where

- $\text{aid}$ is the unique identifier number of a data analysis service.
- $\text{op}$ defines the multiple types of analysis function(s) offered by the data analysis service.
- $\text{ln}$ determines the ability to perform the analysis on preferred human languages.
- $\text{qi}$ determines the set of non-functional properties (e.g., accuracy) of a sentiment analysis service.

\textit{Definition 6:} The integrator service layer first composes the results of multiple data pipelines based on the spatio-temporal requirements and then presents to end user in different visualization formats. A data integrator service $INS$ is a tuple of $<\text{nid}, \text{Pi}, \text{of}, \text{qi}>$

- $\text{nid}$ is the unique service identifier of an integrator service.
- $\text{Pi}$ defines the set of data pipelines aggregated for final composition.
- $\text{of}$ presents the ability to visualize final results in formats (e.g., location maps, summary tables, charts).
- $\text{qi}$ presents the QoS properties (e.g., throughput) of a service.

The composition process of pipelines can be illustrated as follows. 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 services: data collection $DCS_j$, data preprocessing $PRS_j$, location extractor $LES_j$, and data analyzer $DAS_j$. In each service layer, there are $k$ number of candidate services. A row in matrix $M$ represents a data pipeline composed of multiple candidate services from each service layer. Equation 5.2 illustrates the composition of data pipeline $P_j$ as a union product.
of candidate services for a social information service $SN_i$.

$$M = \begin{cases} P_1 = (C_{ci}^{c_1k}, R_{r_1i}^{r_1k}, L_{l_1i}^{l_1k}, A_{a_1i}^{a_1k}) \\ P_2 = (C_{ci}^{c_2k}, R_{r_2i}^{r_2k}, L_{l_2i}^{l_2k}, A_{a_2i}^{a_2k}) \\ \vdots \\ P_j = (C_{ci}^{c_jk}, R_{r_ji}^{r_jk}, L_{l_ji}^{l_1k}, A_{a_ji}^{a_jk}) \end{cases} \quad (5.1)$$

$$P_j(SN_j) = \prod_{i=1}^{k} P_j \cdot [ |C_{ci}^{c_1k}| \cup |R_{r_1i}^{r_1k}| \cup |L_{l_1i}^{l_1k}| \cup |A_{a_1i}^{a_1k}| ] \quad (5.2)$$

Finally, the integrator service layer incorporates the analysis results of all data pipelines into a cohesive format before initializing the visualization of results.

### 5.4.2 Service Composition Layer

The service composition layer constitutes the backbone for the service selection and composition process in the service composition framework. We modify the composition layer with following amendments. First, we incorporate the new service layers and definition structures in the framework. Secondly, the social information service classification model is replaced by QoS model developed in Chapter 4. In addition, the semantic tags of candidate services for each layer are adjusted accordingly. Finally, we develop a new composition approach which is able to assess the QoS features and compose service for spatio-temporal sentiment analysis.

In the remainder of this section, we first present the formal specification of QoS model as a component of framework. Later, we provide the details of graph planning based QoS driven service composition approach.
QoS Model: Social Information Service

In Chapter 4 (Section 4.5.3), we proposed a QoS model of social information services, and defined its different QoS features. In this section, we formalize five of those QoS features used for eSAS service composition. Table 5.1 presents the summary of QoS notations with their definitions.

- **Data Volume**: The data volume $V_N$ determines the quantity of the collected data (e.g., number of tweets, comments) of a social information service $SN_i$. The user of eSAS can put limit on the data items to be collected for the data analysis. The data volume function $f_{\text{Count}}$ counts the data items in a dataset $\Delta$ as following:

\[ V_N(SN_i) = f_{\text{Count}}(\Delta) \]  

- **Data Richness**: The data richness $R_{Sen}$ is the number of unique social sensors present in a dataset $\Delta$ extracted from a $SN_i$. An eSAS user can also specify the minimum number of unique social sensors required in the data analysis. Thus, it can be used to select or reject a dataset prior to analysis. The data richness is calculated as following:

\[ R_{Sen}(SN_i) = f_{\text{SortSen}}(\Delta) \]  

where function $f_{\text{SortSen}}$ sorts the unique social sensors in a given dataset $\Delta$.

- **Data Freshness**: The data freshness $F_R$ determines that the data items in dataset $\Delta$
CHAPTER 5. QOS Driven Service Composition for Big Social Data

are within the temporal bounds. The freshness is defined as below:

\[ F_R(SN_i) = \int_{t_s}^{t_e} \Delta \]  

(5.5)

where \( t_s \) is start time bound, and \( t_e \) is the end time bound for a data item to be included in a dataset \( \Delta \).

- **Data Relevance**: The data relevancy \( D_{Rel} \) establishes that how many data items in dataset \( \Delta \) are relevant to the given sentiment analysis topic. The relevancy of a dataset \( \Delta \) of \( SN_i \) is computed as follows:

\[ D_{Rel}(SN_i) = \frac{A \cap \Delta}{\Delta} \]  

(5.6)

where \( A \) is the set of relevant data items in a given dataset \( \Delta \).

- **Text Type**: The text type \( T_{Type} \) classifies the data items \( d_i \) in a dataset \( \Delta \) into two text types: Informal Text \( T_{Inf} \), Formal Text \( T_{Fo} \). The informal text is written using Internet language (e.g., slang, abbreviations, emoticons), while formal text is written without Internet language. The \( T_{Type} \) is computed as follows:

\[ T_{Type}(\forall d_i \in (\Delta)) = \{d_i : T_{Fo}[T_{Inf}]\} \]  

(5.7)

Quality of Service Driven Composition Approach

The high volume of big social data collected from multiple social information services poses challenge for sentiment analysis. The composition layer requires to analyze QoS features from the collected data and then perform service selection for sentiment analysis. However, as the volume of data grows, it may not be possible to analyze all data from temporary database and estimate some quality features in a limited time-frame. We cope with this problem by using a probabilistic approach for quality feature assessment.

Dynamic QoS Feature Assessment

Let us assume that \( SN_i(D) \) represents a document in the temporary database \( DB \) which contains a finite number of comments as data items \( SN_i(D) = \{d_1, d_2, d_3, ..., d_n\} \) collected from a social information service \( SN_i \). The dynamic feature assessment component first extracts a random subset \( SubSN_i(D) \) of 5% data items for each social information service.
Algorithm 2 Dynamic Feature Assessment

**Input:** Temporary Database $DB$, Social Information Services $SN_i$, Search Terms $K$

**Output:** Data Features Lists $F < SN_i( D_{Rel} , T_{Type}) >$

1. Start Connection ($DB$)
2. for (each $SN_i$ in $DB$) do
3. Extract $SubSN_i(D) \rightarrow (\Delta)$
4. $F < SN_i D_{Rel} > \leftarrow Get D_{Rel} (\Delta, K)$
5. $F < SN_i T_{Type} > \leftarrow Get T_{Type} (\Delta)$
6. Close Connection ($DB$)
7. return $F < SN_i [list : (D_{Rel})(T_{Type})] >$

It then analyzes and evaluates the two QoS features: Data Relevance and Text Type, from the subset. The extracted QoS features are used by composition layer to compose available services. Algorithm 2 formally illustrates the process of feature extraction and assessment.

- **Data Relevance Classifier:** To calculate the relevance $D_{Rel}$ of a dataset, we implement a simple Boolean term frequency $tf$ search model. The term frequency $tf$ search model calculates the frequency of a term (e.g., keyword, expression) occurring in a given data item (i.e., text item). For simple Boolean term frequency model, let us assume that $K = \{k_1, k_2, k_3, ..., k_n\}$ is a set of finite number of keywords used for collecting data from a social information service $SN_i$. The keywords are used to semantically search each data item $d_i$ in subset $SubSN_i(D)$; a $d_i$ is considered a match if at least one keyword is matched. The data relevance $D_{Rel}$ of $SubSN_i(D)$ is computed based on equation 5.6. Finally, the percentage of the relevant data items $P_{Match}(SubSN_i(D))$ of a subset is determined as the probability value of $D_{Rel}$. Generally, if a dataset has lower relevance probability, a probabilistic ranking model based service is used for data filtering; with the higher relevance probability, a simple retrieval model is applied.

- **Text Type Classifier:** For text type classification, we implement a simple binary classifier which labels the data items of a subset $SubSN_i(D)$ into two textual categories: Formal, Informal. The binary classifier is trained with a subset of Internet slang (e.g., lol, omg) taken from Internet slang dictionary\(^4\). After successfully labeling all of the data items of a subset, 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.

\(^4\)https://www.internetslang.com/
Quality of Service Driven Composition Approach

Traditional QoS driven service composition techniques mainly focus on the predefined performance based QoS constraints (e.g., throughput, accuracy) as a service selection criteria. However, in our scenario, the QoS features of social information services are unpredictable, and the composition is primarily dependent on these QoS values. Therefore, a robust composition mechanism is required that can compose services for these changing QoS features.

We have developed a social information service quality features driven service composition approach by using Graph-Planning. The graph-planning allows the development of a set of rules, actions and states for devising dynamic execution plans. We establish QoS attributes as constraints, and define corresponding set of actions and states for composition planning. Thus, the graph-planning would provide the flexibility to service composition layer for assessing the QoS features and then deriving the composition process. The graph-planning approach divides the composition process into two step problem. First, the candidate services are selected based on the data quality features of social information services. Secondly, in case the service composer matches more than one service for a task, performance based QoS properties are used as final service selection. We leverage the existing semantic matching model for service discovery, and traditional performance/QoS attributes (e.g., price, response time) based sorting in the proposed composition approach. The composition approach is described in following three sections.
CHAPTER 5. QOS DRIVEN SERVICE COMPOSITION FOR BIG SOCIAL DATA

Composition Model

The composition model formulates the service composition process as a multi-stage planning problem by using the input of end users. The core concepts to form the Graph-Plan for \( eSAS \) service composition are explained as follows. The Table 5.2 summarizes the notations used in composition approach.

- **Goal:** The goal \( G_{eSAS} \) of planning problem consists of four sub-goals: \( G_{eSaS} = \{G_{DCS}, G_{PRS}, G_{LES}, G_{DAS}\} \), 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 \( PO \). It is comprised of three elements: state transition \( \Sigma \), initial state \( I_S \) and the goal \( G_{eSaS} \), defined as \( PO = \{\Sigma, I_S, G_{eSaS}\} \).

- **State Transition:** \( \Sigma \) consists of three sub-elements: set of states \( S \), set of actions \( Ac \) and set of constraints \( C_s \), defined as \( \Sigma = \{S, Ac, C_s\} \).

- **State:** A 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 filters. For each task, there are \( m \) number of candidate services (i.e., filters) \( T_i = \{ws_1, ws_2, ..., ws_m\} \) available.

- **Action:** \( Ac \) 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, after the completion of data collection \( G_{DCS} \) sub-goal, the data feature assessment task must be completed prior to the execution of data preprocessing \( G_{PRS} \) sub-goal.

- **Task Simulation:** In the \( eSAS \) GraphPlan, task simulation \( F_n(T) \) simulates tasks based on conjunctive actions and constraints as \( F_n(T_i) = \{Ac_i : C_{si}\} \). 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 \( PO \) exists;
Algorithm 3 Graph-Planning Algorithm

**Input:** Initial State $I_S$, Set of constraint $C_s$

**Output:** Composition plan Graph $G$

1. Initialize (Init layer $i$, master table $MT$, graph $G$, init constraint $C_{s0}$)
2. while (all constraints $S_i \rightarrow MT$) do
3. expand graph($G, S_i$)
4. for (each $S_i$) do
5. extract actions and constraints $S_i: (Ac_i : C_{si})$
6. if valid state transition($\Sigma S_i$) then
7. add ($G \leftarrow$ Solution $S$)
8. else
9. abort()
10. $i = i + 1$
11. if (i not changed) then
12. abort()
13. return $G$

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: $G_{cSaaS} \cap T^>_I(\{T_I\}) \neq \{NULL\}$.

Graph-Planning Algorithm

The service composition layer utilizes the composition model to generate the composition plan (i.e., composition graph). The proposed algorithm is extended version of Automated Graph-Planning technique [Siriweera et al., 2017]. In our proposed extension, we limit the algorithm to Forward-Search only to devise a Graph-Plan for find a solution for data pipeline composition. The algorithm 3 is comprised of two basic steps. First, it provides graph planner an initial state $I_S$ 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 composition layer 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 to 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 12 set the termination condition by verifying the solution graph size at the end of each layer processing; if the size of the graph does not change, the algorithm returns failure and aborts the whole composition process for the pipeline. Lastly, line 13 returns the composition graph plan.

**Service Composition Plan Generation**

We use our motivating scenario for Facebook service data pipeline composition graph as an illustration (see Figure 5.3). 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 inputs: the number of social information services required for analysis, temporal and spatial parameters (if any), minimum amount of data volume and richness for the given topic to perform sentiment analysis. $MT$ contains four sets of constraints including pre and post conditions for four sub-goals based on QoS model: Data $D$, Preprocessing $P$, Location $L$, Analysis $A$. For example, 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:
chapter 5. qos driven service composition for big social data

Table 5.3: Summary of Dataset: Collected Data Items

<table>
<thead>
<tr>
<th>Social Information Service</th>
<th>Collected Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter (Tweets)</td>
<td>133043</td>
</tr>
<tr>
<td>Facebook (Comments)</td>
<td>108756</td>
</tr>
<tr>
<td>YouTube (Comments)</td>
<td>3983</td>
</tr>
<tr>
<td><strong>Total Data Items:</strong></td>
<td><strong>245782</strong></td>
</tr>
</tbody>
</table>

data volume \((D : V_N)\) and data richness \((D : R_{Sen})\) are validated, before proceeding 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 \((P : F_R)\), first the irrelevant temporal data is eliminated, then the language filters 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 service does not provide geo-tagged data, thus location extraction task is created for \((LocationServiceSelection \rightarrow G_{LES})\) and state transitions 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 found, the planner aborts the process.

5.5 Experiments and Evaluation

In Chapter 3, we presented the performance of service composition and its feasibility of service composition based sentiment analysis by using a prototype. In this chapter, we evaluate the effectiveness and efficiency of proposed approach. We conduct a set of experiments on the real-world data. We used an example scenario for collecting the approval ratings of American President ‘Donald Trump’ through sentiment analysis. The data is collected between 19-March-2017 to 26-March-2017 from three social information services: Facebook, Twitter and YouTube. Table 5.3 presents the summary of the dataset. For evaluation purposes, we divide our experiment in two parts: 1) Evaluate the performance efficiency of services layers in the composition. 2) Compare overall effectiveness of sentiment analysis results with three traditional sentiment analysis approaches. The prototype is hosted on a local machine, and
5.5.1 Evaluation of Data Collection Service Layer

For each data pipeline composition, we evaluate the processing times of data collection services, and present the corresponding QoS features. We use Throughput as an evaluation measure for calculating the processing times of data collection services and subsequent QoS features evaluation. Throughput $R$ is defined as follows:

$$ R = \frac{d_n}{t} $$  \hspace{1cm} (5.8)

where the $d_n$ is the number of successfully processed items within the unit time $t$.

We calculate the QoS features based on the collected dataset. We simulate the throughput performance of our experiments. For the QoS features evaluation of the dataset, the data volume $V_N$ is simply measured by the data items collected by each data collection service. For the data richness $R_{Sen}$, unique social sensors are sorted based on the social information service generated IDs. Figure 5.4 shows the ratio of unique social sensors with respect to data volume. For calculating throughput, the data collection services: Facebook Graph
API\textsuperscript{5}, Twitter Streaming API\textsuperscript{6} and online YouTube comment scraper\textsuperscript{7}, are invoked three time for collecting datasets with 100, 1000, and 10000, number of data items, and the average throughput of obtaining the data items is calculated. In terms of performance, the Facebook data collection service outperforms other two data collection services for gathering data, while Twitter and YouTube data collection services become almost 2-3 times slower as the amount of data increases. However, the data collection throughput shows that the large data volume can be accessed by trading-off the time. In comparison to the data collection throughput, the sorting of unique social sensor across the datasets remains less than 2.6 seconds for average number of 100,000 data items irrespective of social information service type. The throughput of data collection services of three social information services, and unique social sensor sorting are shown in Figure 5.5 (a-b).

5.5.2 Evaluation of Preprocessing and Location Extraction Service Layer

In this section, for the sake of simplicity, we have combined the performance evaluation of data preprocessing and location extraction service layers. As a first step of data preprocessing service layer, we apply the temporal data filtering on each dataset collected from three social information services. All of the data items which do not comply to the temporal bounds

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.5.png}
\caption{(a)Throughput: Data Collection Services (b) Throughput: Unique Social Sensor Sorting}
\end{figure}
(i.e., data collection dates) are discarded. Figure 5.6 presents the dataset based on the 7 day time-line.

After the data collection, separate random samples of data items are collected for each social information service. Then, the quality features are extracted from each random sample. First, the data relevance $D_{Rel}$ is calculated. The data relevance for Facebook, YouTube and Twitter services is $0.21$, $0.25$, $0.5$, respectively. In order to gain maximum information, we used the ‘Boolean Search Model’ based on inclusion and exclusion technique [Injadat et al., 2016] for all services. The search terms used for collection are considered as inclusion terms. The stop words or phrases are utilized as exclusion terms to remove the irrelevant data items which are not related 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’ is used as a stop word to exclude such data items. For our filtering process, we have used three keywords: ‘US President’, ‘Trump’, ‘Donald Trump’, as inclusion terms. In contrast, we apply three stop words: ‘Ivanka’, ‘Obama’, ‘Hillary’, for discarding the irrelevant data. We assume that the inclusion and exclusion terms are provided by the end users.

Secondly, the text types $T_{Type}$ of the sample datasets are calculated. The classifier computes the **Formal to Informal Ratio** for three social information services. The classification results of formal and informal data classification of three social information services
are presented in Figure 5.7. The random sample of Twitter service has 8.9% of formal data and 91.1% of informal data. In comparison, the random samples of Facebook and YouTube services have 89.2% and 91.8% of formal data, respectively. Only 10.8% and 8.2% of data is classified as informal.

For the location extraction, Facebook and YouTube services do not provide the geo-tagged data for public access. Thus, the geo-tagging condition is determined as False in the composition process. To extract the geo-locations from both Facebook and YouTube services, we used the Stanford NER\(^8\) (Named Entity Recognizer). Although 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. We defined a simple syntactic rule that a location appearing with the conjunction of two words ‘in’ and/or ‘from’ is considered as social sensor’s location. However, more sophisticated semantic methods must be investigated to detect actual locations of social sensors. In contrast, Twitter service provides the geo-tagged data for public access. Hence, the geo-tagging condition for Twitter service is computed as True. For Twitter service, we filtered out all the non-geo-tagged data. The data filters are applied in a linear order such as the output of one filter became the input for next filter. Although a data filtering process may contain a set of filters, each filter can vary in terms of execution.

\(^8\)http://nlp.stanford.edu/software/CRF-NER.shtml
CHAPTER 5. QOS DRIVEN SERVICE COMPOSITION FOR BIG SOCIAL DATA

For instance, the language based filtering turned out to be more time consuming, while location detection and relevant data filtering have better execution time. Figure 5.8 shows the performance time of three data filtering services.

To evaluate the effectiveness of the data preprocessing and location extraction, we apply Signal-to-Noise Ratio (SNR) as an evaluation metric. In each filtering step, Signal is the number of relevant data items remaining after applying the filter, and the Total defines the number of items at the start of the filtering task. For language based filter, we considered the initial number of data items as Total for each dataset.

\[
SNR = \frac{Signal}{Noise} = \frac{Signal}{Total - Signal}
\] (5.9)

Table 5.4 presents the results\(^9\) of SNR for each data filtering step. In the language filtering,

\(^9\)We have combined the results of extra data filter with the language filter
we discard all non-English data items. It can be observed that language based filter has the highest $SNR$. During the irrelevant data filtering, the $SNR$ significantly decreases for inclusion terms based filter. However, the $SNR$ again exceeds 1 for exclusion filter. Finally, the location extraction filter shows the lowest $SNR$ among all of the filters.

5.5.3 Evaluation of Data Analysis Service Layer

For the evaluation of data analysis service layer, we adopt the concept of ‘Black Box Testing’. For sentiment analysis, we use three sentiment analysis services: IBM Watson Natural Language Processing\(^{10}\), Microsoft Text Analytics\(^{12}\), and SentiStrength. Former two services are used for analyzing formal text based on blogs, social information services, and articles which constitute larger text portions. Both services are developed for commercial usage and deployed in cloud settings. The latter is designed to analyze the informal and shorter texts. SentiStrength is an academic project and available for commercial usage.

For evaluation, we only compare the throughput performance of IBM and Microsoft’s services as the SentiStrength is not deployed on any external cloud. We test the services

\(^{10}\)https://www.ibm.com/watson/services/natural-language-understanding/

\(^{11}\)Until 2017, the IBM Watson Natural Language Processing Service was branded as Alchemy API

\(^{12}\)https://azure.microsoft.com/en-au/services/cognitive-services/text-analytics/
performance under the free licence version. Due to the restricted invocations allowed for both services in a limited time, we run the performance test on smaller datasets. The Microsoft service allows the processing of data in batches, where a batch can contain multiple data items. In comparison, the IBM service does not allow batch processing, and analyzes one data item at a time. As a result, the Microsoft service outperforms the IBM service in terms of throughput. Figure 5.9 shows the throughput comparison of both services.

For evaluating the effectiveness of sentiment analysis services, first, a sub-set of dataset is taken from each social information service and manually annotated by human users into three categories: Positive, Negative, and Neutral. For data labeling process, three users are selected who identify English as their first or native language. In the first data labeling step, each user independently classified a data item (e.g., tweet, comment) with a category (e.g., positive). For example, the data item: “@realDonaldTrump you are STUPID” is labeled as negative as it is using foul language or derogatory words. On the other hand, the data item: “I Just got home from watching u Trump I live here in Kentucky we drove up there to support u. Make America Great Again!!!!!” is categorized as positive, since it is written in non-violent and praising language. A data item: “It has only been 65 days of Trumps four year term”, is labeled as neutral, as it does not fulfill the criteria for being positive or negative. As the second data labeling step, the classification of all data items in dataset are consolidated based on the majority classification given by the users for a data item. For instance, if two users labeled a data item as negative, then it is considered as negative.

After completing the labeling process, the same annotated dataset is used for the overall experiments and comparison. Based on the extracted text type from the sample datasets, the data of Facebook and YouTube services is analyzed with IBM service and Microsoft service, respectively. Twitter service data is analyzed with SentiStrength. We used three evaluation metrics: Accuracy $AC$, Precision $PP$, Recall $PR$. The evaluation metric is defined below:

- **Accuracy $AC = \frac{AR}{TR}$**, $TR$ is the number of all data items and $AR$ shows correctly classified data items (e.g., positive).
- **Recall $PR = \frac{PC}{TP}$**, calculates the accuracy for one type of classification (e.g., positive) among the computed data items. $TP$ is the number of all positive data items and $PC$ is the number of correctly classified positive data items.
- **Precision $PP = \frac{PC}{PC+PW}$**, calculates only one type of correct classification (e.g., positive) among the all of correct and wrong results. $PC$ is the number of originally labeled positive data items and $PW$ shows the data items that wrongly classified as positive.
We compare the efficiency of our approach by analyzing the annotated dataset with three different sentiment analysis approaches: Support Vector Machine (SVM), Naive Bayes (NB), and Dictionary Based Approach. The former two approaches (i.e., classifiers) are pre-trained with their baseline datasets\textsuperscript{13}, whereas, the latter is comprised of semantic corpus based on TextBlob [Loria et al.]. It is important to mention here that the three analysis services and their comparison approaches used in evaluation are treated as black boxes.

Table 5.5 provides the evaluation details of three sentiment analysis services including average results of eSAS and comparison approaches. Based on the composition results, the highest accuracy 69.01% is achieved from SentiStrength. On the other hand, Microsoft service demonstrates the lowest accuracy of 63.23%. As SentiStrength is designed specifically for informal text, thus it managed to achieve better results in terms of recall measures and negative precision. Since the Twitter service data contains more negative words and phrases (e.g., swear, abusive language) which are easier to detect, SentiStrength has resulted in 100% negative precision for tweets identification. However, it shows less effective results in terms of positive and neutral precision. Both sentiment analysis services used for Facebook and YouTube services show mixed results with different variations. For instance, both sentiment analysis services exhibit 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 remains higher; whereas the neutral recall remained significantly lower.

Overall the eSAS approach has produced better results in comparison to the three tra-

\textsuperscript{13}http://www.nltk.org/
ditional approaches. For instance, the composite accuracy of eSAS is 66.19%, whereas the second best accuracy results 46.67% are produced by dictionary based approach. In the recall analysis, the dictionary approach produced better results for positive recall 87.5% while eSAS exhibits results better than SVM and NB based approaches. For negative recall, eSAS has demonstrated better results 63.96% than all three techniques. On the other hand, the neutral recall of NB based approach has presented best results 73.68% and the eSAS cultivated results 51.43% better than two services. In the precision analysis, eSAS has established better outcomes for positive and neutral precision by showing 63.75% and 54.84% results respectively. However, for negative precision, the dictionary based approach has slightly better outcome of 75% in contrast to the 71.72% of eSAS which still outperforms the other two techniques.

One reason for variation in results is due to the length of the text (i.e., comments) of both Facebook and YouTube services. 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 lead to significant fluctuation in the output. In contrast, as the models of comparative approaches are not trained and validated with the current datasets, thus their overall performance remains insignificant and mainly exceeds for negative precision and positive recall. The outcome of our comparison points out the limitations of traditional sentiment analysis approaches that they require constant training and validation to acquire better performance output.

5.5.4 Evaluation of Data Integration Service Layer

In this section, we evaluate the performance of the data integrator service layer. The data analysis services used in a data pipeline for extracting sentiment provide results in heterogeneous formats. For example, SentiStrength rates the polarity of negative and positive sentiment at the scale of (±1 to ±5). In comparison, Microsoft Text Analytics API grades the sentiment polarity from (0% to 100%). The score of a data item less than 50% is categorized negative, and the score greater than 50% is classified as positive. Thus, prior to the final integration, the analysis results of each data pipeline must be standardized. We use the following function for the normalizing the sentiment polarity at the scale of 0 to 1.

\[ z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \]  \hspace{1cm} (5.10)

where \( x = (x_1, ..., x_n) \) are the polarity scores of sentiment and \( z_i \) is the \( i \)th normalized value.
After the normalization of the sentiment scores, the final results can be manipulated and visualized by the end users into different formats. We simulated the integration of three data pipelines, where each pipeline contained 100, 1000, 10,000 and 100,000 result outputs. In our simulation, firstly, we created three data pipelines, each producing sentiment results in three formats similar to three sentiment analysis services used in the data analysis layer. Secondly, we performed the normalization of each data pipeline into a cohesive format before the final integration of results for analysis visualization. The data normalization of each data pipeline is repeated 5 times and average time is calculated. The integration layer exhibits the scalable performance around 22.3 seconds for normalizing 100,000 data items of three data pipelines. However, the overall performance of data integration remains dependent on the number of data pipelines, and respective data items per pipeline. Figure 5.10 shows the average throughput time for the integration of three social information services data.

For the data visualization, we implemented a prototype of the eSAS to present the service composition results of sentiment analysis. For first test case scenario, we defined ‘USA’ as a spatial parameter, and random dates as temporal parameters. Figure 5.11 (a) shows the sentiment analysis results by using a live Google map based on social sensor geo-locations. The end user can manipulate the social sensor locations based on different sentiment classifications between given dates. In second test case scenario, the sentiment analysis results are composed as a bar chart. Figure 5.11 (b) presents the overall sentiment analysis results aggregated in a percentage bar chart for three social information services. It
can be observed that most of the social sensors have negative sentiment for US president. Social sensors on Twitter service present highest percentage of negative sentiment, followed by second highest percentage of neutral sentiment. Twitter service has the lowest positive sentiment for the US president. In contrast, social sensors on Facebook and YouTube services have almost equal percentage of negative and positive sentiment. In addition, social sensors on Facebook service present mix opinion with slightly higher degree of negative sentiment. Meanwhile, social sensors on YouTube service has marginally higher positive sentiment for the US president. However, only a limited percentage of social sensors portray neutral sentiment on Facebook and YouTube services.

5.6 Chapter Summary

We developed a novel service composition approach that dynamically composes services for extracting sentiment from big social data generated by various social information services. We formalized and integrated a QoS model for social information services in our existing service composition framework to capture the dynamic features of big social data. We designed a data pipeline composition model which separately processes social information services based on their QoS features. The composition approach dynamically extracts and assesses the quality features of each social information service required for sentiment analysis, and based on the extracted features, it composes required services for sentiment analysis. We formulated a QoS driven service composition technique based on graph-planning. It enables the aggregation of services by utilizing the QoS properties as composition constraints for service
selection. To demonstrate the performance of our proposed approach, we have conducted a set of experiments on real-dataset. In our experiments, we simulated the performance of our composition approach in terms of its execution time. In addition, we compared the effectiveness of our approach. We performed sentiment analysis on test data and compared the results with three traditional sentiment analysis techniques. The results showed the efficiency and effectiveness of our proposed approach.

In the next chapter, we present a composition approach which allows the service composition based sentiment analysis results to be reused.
Chapter 6

Efficient Composition of Reusable Sentiment Analysis Results

The process of extracting sentiment from big social data demands a large amount of computational effort and time. Time-line based sentiment analysis scenarios necessitate the analysis of big social data that was generated between specific periods of time. In such a case, traditional sentiment analysis approaches rely on continuously collecting and storing the big social data before processing it for sentiment analysis based on end users’ requirements. However, the storing and processing cost increases with the volume of the big social data. In Chapter 5, we developed a service composition technique which performs sentiment analysis by using online services that can be aggregated on-demand. Although the service composition technique does not require on-site resources (i.e., infrastructure) like traditional sentiment analysis approaches, the service composition based sentiment analysis process relies on online service provider’s resources. Consequently, as the data size grows, the consumer of online services needs to pay more for utilizing the provider’s services.

In this chapter, we propose a novel meta-information composition approach that makes the service composition based sentiment analysis results reusable. Later, the reusable sentiment results can be composed on-demand and delivered as a service to the end users. The meta-information model includes an information conversion model for storing and accessing the reusable sentiment analysis results as ‘meta-information’. We also devise a meta-information composition algorithm for dynamically aggregating reusable information. We conduct experiments using real-world dataset and the preliminary results demonstrate the feasibility of the proposed model.
CHAPTER 6. EFFICIENT COMPOSITION OF REUSABLE SENTIMENT ANALYSIS RESULTS

6.1 Introduction

Social information services are free online services which continuously produce big social data. In terms of value, the big social data potentially contains a large amount of information (e.g., sentiment) which is subject to various analysis applications. Despite the presence of lucrative and massive information, one major challenge is to efficiently collect, store, process and analyze the staggering amounts of big social data. The distributed technologies such as Hadoop and Spark have been utilized for efficient and parallel processing of large quantities of data. However, these technologies mostly focus on time efficient processing, while at the same time require costly computational resources (e.g., storage, CPU). Consequently, as the size of data grows, the cost of infrastructure also keeps on increasing [Katal et al., 2013; Labrinidis and Jagadish, 2012]. Traditional sentiment analysis and opinion mining applications such as marketing, business oriented customers’ patterns detection and public based political opinion analysis require collection and storage of massive amounts of data over a longer duration. The data is continuously or periodically gathered and stored in application’s repositories, before being extracted and analyzed on-demand for the relevant information. Hence, such long term information analysis applications need the end user to acquire larger on-site storage and processing infrastructure for scalable performance.

In Chapter 5, we developed a service composition approach that dynamically integrates multiple services required for extracting sentiment from big social data. The main advantage of service composition driven sentiment analysis approach is that it does not depend on the computational resources of end users. The resources (e.g., software, hardware) required for service execution are arranged and managed by the service providers. In addition, the end users are not concerned about the underlying complexities of data collection and its processing. However, regardless of such advantages, the services being stateless in nature and do not retain the data or their results after completing their execution. Therefore, for reoccurring analysis scenarios of end users, the services are required to be re-executed. Moreover, during the service execution process, if any of the component service is unavailable, a new composition plan needs to be defined from scratch before the execution. Irrespective of the need for executing an existing service composition plan or devising a new service composition plan, the end users need to pay every time for the service invocation.

The ‘Meta-Data’ provides the information about the data [Singh et al., 2003]. Meta-data contains the information about the creation, transformation, meaning, different types and the quality of the data. The meta-data properties give the ability to effectively manipulate
large datasets in an efficient manner. In contrast to the concept of meta-data, the notion of ‘Meta-Information’ allows the information derived after processing and analysis of actual data to be used [Michener, 2006]. In a typical meta-information system, the required information is extracted from the large amount of data, and stored in meta-information database [Liu and Xiang, 2016]. Afterwards, based on the end users request the meta-information in meta-information database is queried, integrated and delivered on-demand. As a results, the meta-information based systems provide two main advantages: 1) Eliminate the need to store large quantities of the data. 2) Enable the on-demand composition of the required information.

In this chapter, we apply the concept of meta-information modeling for big social data analysis. We develop a meta-information composition technique as an adjunct service layer for our existing service composition framework. The proposed approach transforms the sentiment analysis results obtained from big social data via the eSAS service into reusable information (i.e., meta-information). Later, based on the end users’ requirements, the meta-information is composed and delivered as a service. In comparison to traditional and service composition based sentiment analysis approaches, the proposed approach eliminates the need to collect, store and subsequently process the big social data for reoccurring analysis scenarios. The main focus of this chapter is to formulate a methodology which can efficiently re-utilize the sentiment results which are obtained through various resource consuming processes. Thus, it minimizes the computational and resource consumption costs. The main contributions of this chapter are as follows:

- A meta-information model to transform the service composition based sentiment analysis results into reusable information.
- A composition model to access and deliver the meta-information as a service.

The rest of the chapter is organized as follows. Section 6.2 presents the motivating scenario. Section 6.3 highlights the related work. Section 6.4 provides the solution overview and elaborates the details of the proposed methodology. Section 6.5 provides the details of experimental results and evaluation by real-world datasets. Section 6.6 presents the summary of the chapter.
6.2 Motivating Scenario

We use a public driven political sentiment analysis as our motivating scenario. Let us assume that Sarah is a social information service analyst working for an online news agency Top eNews. After the recent presidential election, the Top eNews agency is interested in monitoring the sentiment of general public from all of the states of country about their newly elected president during his four years term. The agency is keen to analyze that how citizens change their sentiment about their president over that time. Sarah is given the task of developing a system that would enable the agency to analyze the political sentiment of public during that
CHAPTER 6. EFFICIENT COMPOSITION OF REUSABLE SENTIMENT ANALYSIS RESULTS

period. In addition, the desired system must allow the system users to analyze the public sentiment based on the spatio-temporal requirements. Due to limited resources available for the system, Sarah is required to develop the system using limited effort and computational resources (e.g., storage and processing) of agency’s computing infrastructure. Finally, the system needs to deliver the requested information to the users on-demand in different formats over the Internet at any given time.

To avoid the complexities of traditional sentiment analysis approaches, Sarah utilizes the existing service oriented approach (i.e., eSAS) as a baseline solution for extracting sentiment for current scenario. Sarah designs a Web based system with an interface for users that would allow them to get the sentiment analysis results for given spatio-temporal requirements. Sarah utilizes the notion of meta-information modeling technique, and designs a meta-information model that transforms and stores eSAS service results (i.e., meta-information) in a meta-information database. The meta-information service composition layer of the system validates each service request from the system users. For a new analysis request, the composition layer directly invokes the eSAS service that will collect, preprocess and extract sentiments from the social information service data. The results of the eSAS service are first transformed and stored into meta-information database, and then delivered to the user. However, it is possible that the system users may issue an old analysis request for which the eSAS service has already generated the results. In such a case, the eSAS service would be executed again with similar parameters. In order to avoid this repetitive execution of eSAS for reoccurring analysis scenarios, the composition layer composes the information directly from the meta-information database rather than invoking the eSAS, and delivers the results to users. Figure 6.1 presents the overview of meta-information service composition scenario.

The proposed approach has two main benefits. First by avoiding traditional sentiment analysis approaches, Sarah eliminates the need of local computing resources (e.g., storage, processing) and repetitive eSAS executions over the specific time duration (i.e., four years). Secondly, by converting the analysis results into reusable meta-information, she enables the on-demand delivery of information to system users.

6.3 Related Work

Meta-data is usually defined as ‘data about data’. However, meta-data can be defined as a structured description of the essential attributes of an information object [Gill, 2008]. The meta-data can be classified into different types based on the nature of its application. In
earlier research work [Singh et al., 2003], researchers define five generic types of meta-data: 1) User Meta-data allows individual users to associate attributes (e.g., annotations) with data objects or collections. 2) Virtual Organization Meta-data contains the information of data characterizing combination of scientific or collaborative institutions. 3) Domain Specific Meta-data is associated with the ontologies which are developed for specific application communities (e.g., physicists, statisticians). 4) Domain Independent Meta-data properties that describe the data items regardless of their application domain or virtual organization. 5) Physical Meta-data that provides the information of physical characteristics of data on physical storage systems. The meta-data plays a key role for describing, discovering and accessing different datasets in data intensive applications regardless of type [Chervenak et al., 2002]. Meta-data attributes enable the accurate identification of desired data items or multiple datasets for correct analysis and simulation of experiments.

The concepts meta-data and meta-information, have different characteristics. While the meta-data is used to describe and manipulate datasets, the meta-information is mainly used to define the information extracted after processing one or multiple datasets. For instance, the idea of meta-data and meta-information has been applied for various data extensive applications. For instance, in [Alyami et al., 2017], researchers proposed a meta-data based health record management system for organizing and retrieving patients health records during medical emergencies. In another effort [Li et al., 2018], the meta-data analysis of Wi-Fi traffic is used to predict the demographics of online users. In the context of social information service based applications, researchers used the meta-data of Twitter service based social sensors' accounts for modeling their behaviors [Carapinha et al., 2016]. In [Liu and Xiang, 2016], researchers designed a meta-information database that defines the information schema(s) based on user quires, and stores the subsequent extracted information from the Deep Web for later usage. In [Hoiles et al., 2017], investigators analyzed the patterns of the YouTube service based social sensors engagement with different video channels through various types of platform provided meta-data attributes.

In previous chapters, we thoroughly investigated the commercially available online social information service driven information analysis tool and applications. In general, current approaches typically focus on data retrieval task and do not focus on data exploration. Most of these tools generate different types of information analysis results from a limited size of data. In addition, some of these applications lack the flexibility to retain and analyze historical data. Furthermore, as the volume of data exceeds along with different data varieties, it becomes difficult for conventional tools to on-demand visualize meaningful information
[Stieglitz et al., 2018]. To the best of our knowledge, current meta-information systems are mostly designed to utilize the meta-data properties for information retrieval and/or to incentivize the existing information analysis. Furthermore, there is limited research done on how meta-information modeling can be applied to big social data analysis applications (e.g., sentiment analysis).

In contrast to existing approaches, our proposed technique specifically focus on meta-information modeling for the big social data based sentiment analysis. Instead of storing the big social data or repeatedly processing and analyzing it by online services, it transforms the sentiment analysis results into meta-information. The meta-information is subsequently composed and delivered as a service to users making reoccurring or similar requests.

6.4 The Solution Overview: Meta-Information Composition

The proposed meta-information composition approach is divided into three main components: 1) Meta-Information Model 2) Service Request Validator 3) Meta-Information Service Composition. The meta-information model defines the structure and attributes to store a request for invoking the eSAS service, and subsequent results. The service request validator verifies whether a service request is new, old or partially executed. Finally, the meta-Information service composition describes the formal model for meta-information retrieval and integration. The following three subsections provide the details of these three components.

6.4.1 Meta-Information Model

The meta-information model defines the baseline structure for the meta-information database. It is comprised of two type of data models: Analysis Request Data Model, Analysis Result Data Model. The former model keeps the track of user requests, while later records the corresponding results. The data models can be realized by using any database technology including relational databases or document based databases (e.g., XML documents). Table 6.1 summarizes the core concepts of meta-information models. Both of the data models are explained as follows:

Analysis Request Data Model

The analysis request data model saves a request \( R_i \) generated by the user based on its parameters. The analysis parameters are used to invoke the eSAS service. An analysis
Table 6.1: Summary of Meta-Information Models Notations

<table>
<thead>
<tr>
<th>Model Notations</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_i )</td>
<td>An Analysis Request</td>
</tr>
<tr>
<td>( AR_i )</td>
<td>An Analysis Result</td>
</tr>
<tr>
<td>( RID )</td>
<td>Analysis Request Identifier</td>
</tr>
<tr>
<td>( NID )</td>
<td>Analysis Result Identifier</td>
</tr>
<tr>
<td>( SIS_i )</td>
<td>Set of Social Information Services</td>
</tr>
<tr>
<td>( F_i )</td>
<td>Set of Analysis Functions</td>
</tr>
<tr>
<td>( F_{type} )</td>
<td>A Specific Analysis Function</td>
</tr>
<tr>
<td>( A_{value} )</td>
<td>Value Returned by an Analysis Function</td>
</tr>
<tr>
<td>( d_i )</td>
<td>A Data Item</td>
</tr>
<tr>
<td>( T )</td>
<td>Temporal Bounds: ( t_s ) Start Date, ( t_e ) End Date</td>
</tr>
<tr>
<td>( S )</td>
<td>Spatial Requirements</td>
</tr>
<tr>
<td>( Loc )</td>
<td>Geographical Location</td>
</tr>
<tr>
<td>( T_i )</td>
<td>The Time Stamp of a Data Items</td>
</tr>
<tr>
<td>( Vol )</td>
<td>Total Number of Data Items</td>
</tr>
</tbody>
</table>

request is only stored in the database, if the eSAS service is successfully invoked and the sentiment analysis results are received after its execution.

- **Definition 1: Analysis Request.** An analysis request \( R_i \) is defined as a set of five key attributes < \( RID, SIS_i, F_i, T, S \) >, where
  - \( RID \) is the unique identifier of the analysis request.
  - \( SIS_i \) is the social information service to be included in the analysis. For each social information service, the analysis request is further decomposed with same \( RID \) and similar request attributes.
  - \( F_i \) provides the list of required types of analysis functions (e.g., sentiment classification, emotion detection) for a fixed topic of interest.
  - \( T \) defines the temporal bounds \( t_s \) and \( t_e \) (i.e., start date and end date) for which the sentiment analysis is required.
  - \( S \) determines the spatial requirements of user such as country or states for the analysis.

**Analysis Result Data Model**

The analysis result data model defines the structure to save the sentiment analysis results obtained from the eSAS service execution. The analysis results are separately stored for each social information service against a request. For instance, if an analysis request included
two social information services (e.g., Facebook and Twitter), then two separate records are created and stored in the meta-information database. Figure 6.2 presents the relationship of both models with respect to an analysis request.

- **Definition 2: Analysis Result.** An analysis result $AR_i$ of $eSAS$ service for a social information service is a set of attributes $< NID, RID, SIS, F_{type}, A_{value}, Loc, T_i, Vol >$, where
  
  - $NID$ is the unique identifier of the analysis result.
  - $RID$ is the unique identifier to trace the original analysis request.
  - $SIS$ is the targeted social information service in the analysis request.
  - $F_{type}$ is the specific analysis function (e.g., sentiment classification) for which the $eSAS$ service has performed the analysis.
  - $A_{value}$ is the functional value returned by the requested analysis for a data item $d_i$. For instance, for sentiment classification function a value can be a sentiment classification such as positive, negative, or neutral.
  - $Loc$ defines the geographical locations of social sensors with respect the analysis value.
CHAPTER 6. EFFICIENT COMPOSITION OF REUSABLE SENTIMENT ANALYSIS RESULTS

- $T_i$ defines the time stamp (i.e., date) of a data item.
- $Vol$ determines the number of data items analyzed.

6.4.2 Service Request Validator

The meta-information service composition layer receives an analysis request, and forwards it to the validator component before invoking the $eSAS$ service. The validator verifies the execution status of all requests from existing analysis request records to make sure that the $eSAS$ service is not invoked for similar requirements. The validator verifies each request for four parameter: $< SIS_i, F_i, T, S >$, for the following three scenarios:

- **Existing Request:** A request $R_i$ is considered as an existing request based on two scenarios. In the first scenario, if the all parameters of a request are matched with a single record, it is considered as an existing request. In the second scenario, a request is considered an existing one, if the parameters of request are matched within multiple records. For instance, let us assume that there are two existing requests: $R_1 :< SIS_i, F_i, T : (t_1 - t_2), S >$, $R_2 :< SIS_i, F_i, T : (t_3 - t_4), S >$, where $T$ contains the start and end dates. For a current request $R_3 :< SIS_i, F_i, T : (t_2 - t_3), S >$, the validator selects all of the records which match the three parameters $< SIS_i, F_i, S >$. Next, it defines a time-line of retrieved records as a composite record $R_1 + R_2 :< SIS_i, F_i, T : (t_1 - t_4), S >$. The $R_3$ is considered as an existing request as three parameters $< SIS_i, F_i, S >$ falls within the range of temporal bounds $T : (t_2 - t_3)$. The service layer later uses the $RID$s of $R_1$ and $R_2$ with temporal bounds $T : (t_2 - t_3)$ to extract the sentiment analysis results.

- **Partial New Request:** A request $R_i$ is determined as partially new, if the parameters $< SIS_i, F_i, S >$ are complete match and temporal bound are partially matched. For instance, let us assume that there are two existing requests: $R_4 :< SIS_i, F_i, T : (t_5 - t_6), S >$, $R_5 :< SIS_i, F_i, T : (t_7 - t_8), S >$. For a current request $R_6 :< SIS_i, F_i, T : (t_4 - t_{10}), S >$, the validator selects all of the records which match the three parameters $< SIS_i, F_i, S >$ and establish a time-line as a composite record $R_4 + R_5 :< SIS_i, F_i, T : (t_5 - t_8), S >$. The $R_6$ is determined as a partial match as it is within the temporal range of $T : (t_5 - t_8)$. The validator adjusts the request $R_6$ for missing parameters as $R_6 :< SIS_i, F_i, T : (t_9 - t_{10}), S >$. It returns the $RID$s of $R_4$ and $R_5$ to service layer and $R_6$ to be invoked as a $eSAS$ service for the partially new parameters.
CHAPTER 6. EFFICIENT COMPOSITION OF REUSABLE SENTIMENT ANALYSIS
RESULTS

- **New Request:** An analysis request is conceived as a new request if all of its parameters 
  \(< SIS_i, F_i, T, S >\) are not exactly matched with any individual service request records.
  In addition, they do not match within the bounds of multiple service request records.

### 6.4.3 Meta-Information Service Composition

The meta-information service composition layer is responsible for aggregating the meta-
information and presenting it to the users. After the validation of an analysis request, the 
response of validator is received by the composition layer. Based on the outcome of service 
request validator, the composition layer may perform a combination of the following tasks.

- **Meta-Information Creation:** The service request validator returns a request \( R_i \) 
as a new or partially new request. In either case, by using the provided parameters, 
the \( eSAS \) service is invoked. The following function \( F_{Invoke} \) defines the \( eSAS \) service 
invocation for request \( R_i \):

\[
F_{Invoke}(R_i) = eSAS : R_i
\]  

(6.1)

After receiving the response from the \( eSAS \) service, the analysis results are stored in 
meta-information database \( MIDB \). In case of partial new request, the analysis request 
records are updated in \( MIDB \).

- **Meta-Information Retrieval:** In case, a service request is identified as an exist-
ing request, the validator provides the composition layer a list comprised of analysis 
request ID(s) \( RID \). The composition layer retrieves the analysis records from the 
meta-information database based on the given \( RID \)s. The following function \( F_{Retrieve} \) 
presents the meta-information retrieval process.

\[
F_{Retrieve}(List[RID_i]) = \sum[AR_i \in RID_i] \cap (MIDB)]
\]  

(6.2)

In addition, if a service request is considered as a partial new request, the service 
request validator returns the \( RID \)s of existing records. Similarly, the above function is 
utilized for analysis record retrieval.

For a partial new analysis request, the composition layer does not start the retrieval process 
until the \( eSAS \) service successfully provides the sentiment analysis results for the missing 
parameters. After receiving the \( eSAS \) service response, the composition layer aggregates the
CHAPTER 6. EFFICIENT COMPOSITION OF REUSABLE SENTIMENT ANALYSIS

RESULTS

complete results and presents in different formats. The complete process of meta-information service composition is illustrated in Algorithm 4.

**Algorithm 4** Meta-Information Service Composition

**Input:** Analysis Request $R_i$, Meta-Information Database $MIDB$

**Output:** Analysis Result $AR_i$

1: Start Connection (MIDB)
2: $VRes = ServiceRequestValidator(R_i)$
3: if $VRes == $ Existing Request then
4: $GetRIDs (VRes) → List < RID >$
5: $F_{Retrieve}(List < RID >) → AR_i$
6: if $VRes == $ Partial New Request then
7: $F_{Invoke}(R_i) → AR_i$
8: $Save(AR_i → MIDB)$
9: $GetRIDs (VRes) → List < RID >$
10: $F_{Retrieve}(List < RID >) → AR_i$
11: if $VRes == $ New Request then
12: $F_{Invoke}(R_i) → AR_i$
13: $Save(AR_i → MIDB)$
14: Close Connection (MIDB)
15: return $AR_i$

6.5 Experiments and Evaluation

To evaluate the proposed approach, we implemented a prototype by utilizing our motivating scenario with different test cases. We performed the experiments in two folds: 1) We collected the data from two social information services: Facebook and Twitter, and extracted the sentiment analysis results. We implemented the meta-information model to transform and store the sentiment analysis results. 2) We compared the meta-information composition performance with the eSAS service execution time for reoccurring scenarios. The experiments are conducted using Microsoft Visual Studio 2017 .Net framework with ASP.Net/C# on a 3.40 GHZ Core i7 processor and 8 GB RAM by using Windows 7 as operating system. We implement the meta-information database by using built in Microsoft SQL Database of Net framework. The data from Facebook service and Twitter service is collected for seven days as ‘President Donald Trump’ the topic of interest. Table 6.2 provides the details of the collected dataset.
CHAPTER 6. EFFICIENT COMPOSITION OF REUSABLE SENTIMENT ANALYSIS RESULTS

Table 6.2: Summary of Dataset

<table>
<thead>
<tr>
<th></th>
<th>Facebook (Posts)</th>
<th>Twitter (Tweets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1</td>
<td>5606</td>
<td>12165</td>
</tr>
<tr>
<td>Day 2</td>
<td>11367</td>
<td>20025</td>
</tr>
<tr>
<td>Day 3</td>
<td>10840</td>
<td>20117</td>
</tr>
<tr>
<td>Day 4</td>
<td>10023</td>
<td>20098</td>
</tr>
<tr>
<td>Day 5</td>
<td>29242</td>
<td>20124</td>
</tr>
<tr>
<td>Day 6</td>
<td>26644</td>
<td>20361</td>
</tr>
<tr>
<td>Day 7</td>
<td>15034</td>
<td>20153</td>
</tr>
<tr>
<td>Total</td>
<td>108756</td>
<td>133043</td>
</tr>
</tbody>
</table>

6.5.1 Meta-Information Model Implementation

For the meta-information model evaluation, we first implemented both of the meta-information models as two tables of meta-information database. We modeled the motivating scenario as following test case. A user ‘X’ requests to analysis the sentiment of US citizens for their president from two social information services: Facebook and Twitter. The same request is repeated for seven consecutive days. Second, we assume the sentiment analysis results are provided by the eSAS service. We simulate the performance of transforming and storing of the analysis results into meta-information database.

We generate the analysis request and analysis response structures based on the meta-information models. For visualization, we used the XML based documents to present the requests and responses. However, the meta-information is stored in the relational database. The test case analysis request for day 1 is factorized as follows. A system generated unique identifier \( RID \) to identify the request, two social information services as \((SIS_1)\) Facebook and \((SIS_2)\) Twitter, analysis function \( F \) as sentiment classification, requested analysis date as \( T \), and ‘US’ as a spatial requirement \( S \). The above analysis request is defined in XML structure in Figure 6.3. For the same analysis request, the response of the eSAS service is mapped into analysis result as following. The analysis results are first separated by each social information service \( SIS \). In each group, the data items \( d_i \) are further sorted by the functional value \( A_{value} \) of the requested sentiment analysis function \( F_{type} \). For a given functional value (e.g., positive, negative) the locations are grouped by the unique geo-locations with date \( T_i \). Finally, a tuple of record is created with an unique identifier \( NID \) and the analysis request identifier \( RID \) which has the sentiment analysis function, number of data items \( Vol \) originating from a location categorized with the functional value. Figure 6.4 presents the analysis response as XML document.
In the next step, we evaluated the performance of transforming the eSAS service data as meta-information. We assume that the response of eSAS is received by the meta-information service composition layer in a standardized format (e.g., CSV). We simulate the throughput of meta-information transformation for three social information services. As a part of experiment setup, we run the transformation process on a unified dataset (i.e., eSAS response). In first iteration, we execute the transformation for one social information service. For second iteration, we perform the transformation for two social information services. In third iteration, we use three social information services for transformation. In each run, the dataset contains 100, 1000, 100,000, and 100,000 data items. The service composition layer performs the grouping and subsequent conversion of results into meta-information structure. The composition layer shows the scalable performance as the execution time increases in linear order. For instance, the performance remains comparatively consistent for three iter-
Chapter 6. Efficient Composition of Reusable Sentiment Analysis Results

Figure 6.5: Meta-Information Transformation Time

Ations up to 100,000 data items. However, the execution time differences become significant as the number of social information services increases (i.e., 1 service 5.02 seconds, 2 services 6.2 seconds and 3 services 11.2 seconds). Regardless of such deviation in execution time, the overall performance remained scalable. Figure 6.5 shows the average throughput time of the meta-information transformation.

6.5.2 Meta-Information Composition Performance Evaluation

In this section, we evaluated the performance of meta-information service composition. After transforming and storing the analysis results into meta-information database, the composition layer can directly utilize and aggregate the meta-information for reoccurring requests. To demonstrate the efficiency and applicability of the meta-information modeling, we compared the performance of meta-information composition with the service composition based sentiment analysis approach. We designed the test case scenario as follows. First, based on the test case scenario of previous section, the sentiment analysis results are extracted from the dataset by using the eSAS service for each day, and saved into meta-information database. Second, we assumed that a different user ‘Y’ repeats the same analysis request with similar parameters. For the new user, the analysis results are composed from meta-information.
database. We compare the execution time of both approaches.

For the eSAS service based sentiment analysis results, we computed the combined throughput time of data collection services, data preprocessing services, location extraction services, and data analysis services for each day. For this experiment, we employed Microsoft Text Analytics service for the data analysis. To allow comparison, we executed the meta-information composition requests for each day in three parts. In the first fold, the meta-information composition is executed a single time. In second and third folds, the combined information composition throughput is computed for 10 and 100 iterations. Figure 6.6 shows the comparison results of the eSAS service and meta-information service compositions throughput. Based on the comparison, the meta-information composition clearly outperforms eSAS. As an analysis request which is executed 100 times has the maximum combined throughput of almost 0.05 seconds in contrast to 280 seconds of the eSAS service. Thus, it reveals the performance for repeating scenarios with meta-information modeling which stores sentiment analysis results. In addition, it demonstrates that the meta-information modeling would substantially reduce the processing time and resource consumption (i.e., cloud service invocations).
CHAPTER 6. EFFICIENT COMPOSITION OF REUSABLE SENTIMENT ANALYSIS
RESULTS

6.6 Conclusion

This chapter proposed an efficient meta-information modeling technique that transforms and stores sentiment analysis results of big social data as reusable information (i.e., meta-information). Later, the meta-information is composed based on the users request, and delivered as a service. We developed a meta-information model which keeps the records of sentiment analysis requests, and stores the subsequent analysis results as meta-information. As a result, for reoccurring sentiment analysis requests, the resource consuming sentiment analysis tasks (e.g., data collection, preprocessing, analysis) are not repeated, instead the meta-information as analysis results are aggregated and presented to the users. We devised an algorithm for composing meta-information as a service. We implemented a prototype as a proof of concept. We conducted experiments on the real-world data. In our experiments, we demonstrated the efficiency of our approach by comparing the performance scalability with the service composition based sentiment analysis approach. The preliminary experiment results show the applicability our approach in terms of execution time.
Chapter 7

Conclusion

Social information services have become a free source of social sensor data which contains a plethora of subjective information (e.g., sentiments, opinions). One common mechanism to extract subjective information is sentiment analysis. However, traditional sentiment analysis approaches require specialized skills and rely on laborious and time-consuming tasks like datasets cleansing, data labelling, data model training and validation. In this thesis, we have devised a novel service-oriented framework for social information services based sentiment analysis that provides a substitute solution option for traditional sentiment analysis approaches. Our approach utilized services as a mean to perform various sentiment analysis tasks while hiding the underlying complexities of traditional approaches. In the following sections, we summarize the outcomes of our research questions and key contributions of this thesis. Finally, we discuss several potential future research directions.

7.1 Summary of Contributions

In this section, we present a summary of the proposed methodologies and their findings based on our four research questions.

Designing A Service Composition Framework For Sentiment Analysis

In Chapter 3, we designed a service composition framework that is capable of composing services for social information services based spatio-temporal sentiment analysis. In this contribution, our main aim was to devise an alternative solution approach for existing labor intensive data centric sentiment analysis methods. Existing sentiment analysis approaches use multiple complex processes (e.g., data collection, preprocessing, information extraction),
whereas these processes require special skills and manual human effort. By formalizing the sentiment analysis problem as a service composition solution, third party services were used to replace and perform the typical sentiment analysis tasks. Thus, our approach provides greater flexibility, automation and abstraction for overall complexities of sentiment analysis applications where spatial and temporal information is highly desired. In our framework, we first developed a service based architecture which defines a set of service layers and their component services which depict the sentiment analysis processes such as data collection, preprocessing (e.g., noise filtering), location extraction and sentiment extraction. We presented the models of different service layers, component services with varying features. Secondly, we devised a service composition technique to aggregate component services from different service layers. For service composition, we formulated a classification model to categorize social information services based on their generic features. Next, we designed a human centric tagging model which uses the classification model’s properties to semantically annotate the candidate component services. Finally, we developed a composition algorithm that retrieves the component service by matching the semantic tags, and composes appropriate services required for data collection, preprocessing, location extraction and sentiment analysis. For evaluation, the experiments were conducted by using a flu surveillance case study based on social sensors sentiments, and the results demonstrated the applicability of our framework.

Recognition of Social Information Services Features For Different Domains

The data produced by social information services have changing features with respect to different domains or topics of interest such as politics, entertainment and health. One challenge is the dynamic identification and appropriation of changing features. In Chapter 4, we applied the theory of ‘Service-Orientation’ on social information services and hypothesized them as conventional cloud/Web enabled services. By applying the notion of service-orientation, our main goal was to investigate the functional and non-functional (i.e., QoS) features of social information services, and their variations with respect to different domains. Traditional sentiment analysis approaches treat social information services as homogeneous data source rather than as typical services. In addition, these approaches assume that social information services (e.g., Facebook, Twitter) have similar data characteristics. The functional and specifically QoS features of social information services can be utilized for finding and composing appropriate processing and analysis services for sentiment analysis related tasks. To show the validity of our hypothesis, we first defined an abstract service model
to present social information services as cloud services. Secondly, we developed two models to present the functional and QoS features of social information services. The Ontology Web Language for Service (OWL-S) was used to illustrate a Web accessible service model for social information services. We conducted a series of experiments on the data obtained from multiple social information services for different domains. The results have shown that service-oriented modeling is a viable option to depict changing features of social information services.

**Dynamic QoS Driven Service Composition For Big Social Data Analysis**

The extraction and integration of sentiments resulting from ever growing volume of big social data is a challenging task. The overall sentiment analysis process becomes more complicated due to the dissimilar features of big social data originating from different social information services. Considering these challenges, in Chapter 5, we developed a dynamic QoS-aware service selection and composition approach for extracting sentiment analysis from big social data. We incorporated the QoS model developed in Chapter 4 with our new service composition approach. The new approach is able to simultaneously compose services for extracting, processing and analyzing big social data of individual social information services for sentiment analysis based on their unique and changing QoS features. We addressed three main challenges in our framework enhancement. First, we expanded the existing service architecture which can concurrently process, analyze and integrate sentiments obtained from a number of social information services. Secondly, we developed a QoS feature assessment model which allows to dynamically extract and assess the QoS features of any given social information services based on their data. Finally, we devised a new service composition algorithm based on Graph-Planning which uses the extracted QoS features for devising the service composition plan. We conducted experiments and compared our approach with existing sentiment analysis techniques to show the effectiveness of our approach. We demonstrated that our composition technique is scalable and can be implemented in real life settings for big social data analysis scenarios.

**Efficient Reuse And Composition of Sentiment Analysis Information**

The service composition based sentiment analysis of big social data is a time and resource consuming task. The results produced by costly big social data processing required to be efficiently re-utilized. In Chapter 6, we developed a new meta-information modeling mechanism
that enabled the reuse of sentiment analysis results obtained from big social data. We applied meta-information modeling approach to transform and preserve the sentiment analysis results as reusable information (i.e., meta-information) in meta-information database. Later, the meta-information is extracted from meta-information database and delivered as an on-demand service. The proposed approach can be applied for re-utilizing information extracted from resource intensive applications, as requests to process and analyze large amounts of big social data are often repeated. In this contribution, we first devised two meta-information models. The first meta-information model tracks the records of sentiment analysis service composition requests, while the second model converts and stores the results into reusable meta-information structure. Secondly, we devised a meta-information composition algorithm that aggregates the required sentiment analysis information from meta-information repository. Based on our experiment on real datasets, the results proved that the meta-information modeling and composition approach is a feasible and efficient solution for sentiment analysis applications as it helps to avoid repetitive and large scale data processing.

7.2 Future Research Directions

Social information service analysis is an emerging area of research. Combining it with research domains such as sentiment analysis provides the ability to develop powerful social information analysis applications. Moreover, the inclusion of cloud and services computing paradigm with social information service based information analysis has opened the door for various research opportunities. In the following, we highlight some promising research directions.

Designing Efficient Services For Big Social Data Collection and Processing

The growing number of social sensors on social information services is resulting in explosion of big social data. This sheer amount of data brings various opportunities and challenges in terms of data collection, preprocessing and information extraction. For instance, current data collection services such as platform provided APIs and third party tools (e.g., scraper) put several restrictions (e.g., limited data size, formats) for data gathering. Thus, these services may not be able to cope with the increasing volume of data, and the demand for real-time delivery for various analysis applications. In addition to the limitations of current data collection services, obtaining relevant data for specialized applications is a complex task. Generally, after the data collection, various data preprocessing steps such as noise removal, relevant data filtering, and data formatting are required before the data can be
be analyzed. These data preprocessing tasks are complex and extensively time-consuming. Therefore, there is an urgent need for scalable and efficient ‘Data Services’ which can on-demand deliver big social data without any restrictions which is ready to be consumed for information extraction.

Enhancing the QoS Models of Social Information Services

With the increasing popularity of social information services, more and more social sensors across the globe are utilizing these services and sharing their data. Social sensors are data generating sensors which have their own various features such as age, gender, ethnicity, religious and political affiliations. The features of social sensors may influence the behaviors (i.e., functional and non-functional features) and usage patterns of social information services. The social sensor specific features can help to break down any information extracted from social information service at fine grained levels. For instance, with the help of social sensor specific features the sentiment analysis results can be visualized based on gender, age groups and multiple affiliations. Currently, limited research has been done studying the relationship of social sensor specific features on social information services data. In this thesis, we only investigated the QoS features of social information services. Therefore, in order to develop sophisticated social information analysis applications, it is important to develop new QoS models which can specify the composition of social sensors information based on their different features.

Social Information Services Integration with IoT Services

The Internet of Things (IoT) is a system of interrelated computing sensors, physical devices, and objects which produce and share data over the network without the interaction of human users [Gubbi et al., 2013]. The IoT sensors generate an enormous amount of data which is delivered as a service (i.e., IoT services) to different application such as public health and safety, disaster management, smart transportation and security surveillance. The main difference between social information services data and IoT data is that the former contains the subjective data (e.g., sentiment) of social sensors, while the latter contains the application specific data. However, there are many applications and scenarios which can benefit from combining both types of data. For instance, social information services may provide data for disaster monitoring from locations where IoT sensors are not present or fail to deliver the required data, and vice versa. In such cases, either type of sensor data can be used as a
replacement for the missing data as well as compliment each other. Hence, there is a need for developing tools, middlewares and frameworks which can efficiently process and integrate social information services with IoT services for different data-centric applications.

**Developing the Big Social Information Analysis Services**

The growing volume of big social data has opened new challenges and venues within the research domain of big data. Moreover, the high demand of big social data in multiple applications shows that it contains the multi-dimensional information. Thus, one main issue is to minimize the time, effort and cost while transforming the big social data into multiple types of information with maximum re-usability. In this thesis, we developed a service oriented approach for big social data based sentiment analysis. We devised a meta-information modeling approach which transforms the sentiment analysis results into re-usable information. However, various types of information analysis require development of different sets of service oriented strategies and techniques to convert big social data into big social information. Although there are several approaches developed for parallel and time-efficient data processing, devising the techniques of re-usable information remains largely unexplored. Thus, there is an opportunity to develop ‘smart’ techniques which can be applied for information re-usability.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


T. Poell and E. Borra. Twitter, youtube, and flickr as platforms of alternative journalism: The social media account of the 2010 toronto g20 protests. *Journalism*, 13(6):695–713, 2012.


BIBLIOGRAPHY


