Twitter mining for ontology-based domain discovery incorporating machine learning

Bilal Abu-Salih, Pornpit Wongthongtham and Chan Yan Kit

Abstract

Purpose – *This paper aims to obtain the domain of the textual content generated by users of online social network (OSN) platforms. Understanding a users' domain (s) of interest is a significant step towards addressing their domain-based trustworthiness through an accurate understanding of their content in their OSNs.*

Design/methodology/approach – *This study uses a Twitter mining approach for domain-based classification of users and their textual content. The proposed approach incorporates machine learning modules. The approach comprises two analysis phases: the time-aware semantic analysis of users' historical content incorporating five commonly used machine learning classifiers. This framework classifies users into two main categories: politics-related and non-politics-related categories. In the second stage, the likelihood predictions obtained in the first phase will be used to predict the domain of future users' tweets.*

Findings – *Experiments have been conducted to validate the mechanism proposed in the study framework, further supported by the excellent performance of the harnessed evaluation metrics. The experiments conducted verify the applicability of the framework to an effective domain-based classification for Twitter users and their content, as evident in the outstanding results of several performance evaluation metrics.*

Research limitations/implications – *This study is limited to an on/off domain classification for content of OSNs. Hence, we have selected a politics domain because of Twitter's popularity as an opulent source of political deliberations. Such data abundance facilitates data aggregation and improves the results of the data analysis. Furthermore, the currently implemented machine learning approaches assume that uncertainty and incompleteness do not affect the accuracy of the Twitter classification. In fact, data uncertainty and incompleteness may exist. In the future, the authors will formulate the data uncertainty and incompleteness into fuzzy numbers which can be used to address imprecise, uncertain and vague data.*

Practical implications – *This study proposes a practical framework comprising significant implications for a variety of business-related applications, such as the voice of customer/voice of market, recommendation systems, the discovery of domain-based influencers and opinion mining through tracking and simulation. In particular, the factual grasp of the domains of interest extracted at the user level or post level enhances the customer-to-business engagement. This contributes to an accurate analysis of customer reviews and opinions to improve brand loyalty, customer service, etc.*

Originality/value – *This paper fills a gap in the existing literature by presenting a consolidated framework for Twitter mining that aims to uncover the deficiency of the current state-of-the-art approaches to topic distillation and domain discovery. The overall approach is promising in the fortification of Twitter mining towards a better understanding of users' domains of interest.*

Keywords *Ontology, Machine learning, Twitter mining, Domain discovery, Domain-based trustworthiness*

Paper type *Research paper*

Bilal Abu-Salih and Pornpit Wongthongtham are both based at Curtin University, Perth, Australia.
Chan Yan Kit is based at Department of Electrical and Computer Engineering, Curtin University, Perth, Australia.

1. Introduction

The demand for real-time business intelligence and the popularity of social media have created a need for social business intelligence. Social business intelligence aims to reveal the fundamental factors derived from social perspectives, which determine an

Received 12 November 2016
Revised 26 September 2017
Accepted 1 October 2017

organisation's performance. People express their thoughts, feelings, activities and plans via online social networks (OSNs). Often, their posts link to product (s), service (s), event (s), society or person(s) and people in OSNs intuitively tend to seek and connect with like-minded people. This homophily results in building homogenous personal networks based on behaviours, interests and feelings (McPherson *et al.*, 2001). The rapid increase in unstructured social data has highlighted its importance as a means of acquiring deeper and more accurate insights into businesses and customers. In particular, OSNs are a medium for content makers to express and share their thoughts, beliefs and domains of interest. This gives individuals access to a wider audience which positively affects their social rank and provides other benefits, such as gaining political support (Rainie and Wellman, 2012). Therefore, the cornerstone of building users' online social profiles is a veritable understanding of their domains of interest.

Because of the open environment and limited restrictions of social media, rumours can spread quickly and false information can be broadcast rapidly. This may have adverse effects on businesses, political management and public health, particularly if the false information is being published together with trustworthy information. However, if it is an accurate information, this could be greatly beneficial to individuals and organisations as a means of acquiring value from social media data. Spam is a well-known category of low-quality content. Social spam content such as fake accounts, bulk messaging (sending the same post many times in a relatively short period of time), malicious links and fake reviews lower the quality of experience of social community members (Lee *et al.*, 2010). Social media data are big, heterogeneous and unstructured in its textual content, structured in its metadata, can be linked and have different trust levels. Sherchan *et al.* (2013) defined "trust" as the measurement of confidence that a group of individuals or communities will behave in a predictable way. Trust in social media refers to the credibility of users and their shared content in a particular domain. Users are known to be trustworthy in a particular domain. However, this does not mean that their trustworthiness will have the same value in other domains. The trustworthiness of social media data is now crucial (Abu-Salih *et al.*, 2015). With such a vast volume of data interchanged within social media ecosystems, determining domain-based data credibility is considered a vital issue. The importance of domain-based trust in the social media context originates from the affluent resources for market analysis, e.g. the voice of customer (VoC) and the voice of market (VoM), recommendation systems, domain-based influencers' discovery and the like. Hence, understanding users' domain (s) of interest is a significant step in addressing their domain-based trustworthiness through an accurate understanding of their content temporally in OSNs.

In this context, companies incorporate advanced social data analytics when designing effective marketing strategies and seek to leverage the interactive quality of OSNs. Thus, to create the required interaction with their customers, companies use many modern communication to attract customers and visitors to their online social platforms. Consequently, it is necessary for companies to analyse their customers' social content and classify the customers into appropriate categories on the basis of their topics of interest, to deliver the right message to the right category.

Most of the existing approaches to this topic rely on bag-of-words techniques, such as latent Dirichlet allocation (LDA) (Blei *et al.*, 2003). However, despite the importance and popularity of these techniques for inferring the users' topics of interest, when it comes to the use of Twitter, there are three main shortcomings of such an approach:

1. the inability to consider the semantic relationships of the terms in the user's textual content;
2. the inadequacy of its application to a topic modelling technique using short text messages, such as tweets; and

3. the high-level topic classifications that use these bag-of-words statistical techniques are inadequate and inferior (Michelson and Macskassy, 2010).

On the other hand, incorporating semantic Web consolidated tools such as AlchemyAPI™ [1], offers a comprehensive list of taxonomies divided into hierarchies, where the high-level taxonomy represents the high-level domain and the deeper-level taxonomy provides a fine-grained domain analysis. For instance, "art and entertainment" is considered a high-level taxonomy in which "graphic design" is one of its deep-level taxonomies. LDA is unable to provide high-level topics such as "art and entertainment" from a corpus of tweets unless this term exists in the corpus. Semantic analysis, conversely, extracts semantic concepts and infers high-level domains through analysing the semantic hierarchy of each topic, leveraging an ontology; this is not possible when using an LDA technique.

The main challenge in obtaining the accurate domain of a tweet is the ability to accurately determine the classification of its textual content. This is because of the several features of linguistics, such as polysemy (where the same word has several meanings), homonymy (where words have the same spelling and pronunciation, but have different meanings) and contronymy (where the same word has contradictory meanings). This diversity in linguistics makes the process of determining the correct domain of interests from the short textual content of the tweet more difficult. Hence, it is essential to obtain an accurate understanding of the semantics of the tweet text to determine the user's domain of knowledge. This will assist in determining the topic/domain of the tweets that will be posted by the user in future. This paper aims to address this problem by proposing a comprehensive framework incorporating semantic analysis and machine learning.

Semantic analysis, through existing ontologies and linked data, enables the eliciting of knowledge from social data, thereby enriching its textual content to deliver semantics, and links each message with a particular domain. Machine learning applications enable real-time predictions leveraging high-quality and well-proven learning algorithms. On the basis of the current dominant position and high impact on business in several used cases, according to Gartner's recent report on emerging technologies[2], incorporating machine learning in particular enhances the decision-making process and provides valuable insights from large-scale data.

This study presents an approach to glean profound insights into users' domains of interest from their pervasive propagation of tweets. This is achieved through a systematic approach beginning by addressing the volume quality of social big data incorporating data generation and acquisition techniques, and then inferring the added value obtained from the data analysis. This aims to contribute to an advanced domain-based trustworthiness approach that is able to filter out unsolicited tweets and increase the value of content. To achieve this objective, this paper presents a consolidated framework leveraging former knowledge obtained from an analysis of the user's historical content. In this context, the politics domain is used to determine the user's interest in this domain. Hence, we propose an effective approach to classify Twitter users and their new updates according to two main categories:

1. *On-topic*: a user or tweet is classified under the politics domain.
2. *Off-topic*: a user or tweet is classified under the non-politics domain.

The proposed approach comprises two main analysis phases incorporating several semantic analysis tools and machine learning modules. In the first phase, the users' historical tweets are collected and their interest is examined over time thereby providing a prediction of the users' interest, taking the temporal factor into consideration. In the second phase, the outcome of the previous analysis is used as a primary input to forecast the domain of future tweet content. Users' classification is achieved through the use of well-known machine learning classifiers. A comparison is conducted to benchmark the performance of the incorporated machine learning modules.

The main contributions of this paper are summarised as follows:

- A time-aware framework incorporating comprehensive knowledge discovery tools and well-known machine learning algorithms is proposed for domain-based discovery, which is applicable to the Twittersphere platform and customisable to other OSNs.
- The proposed framework is able to perform classification tasks at the user level and tweet level.
- The conducted experiments verify the effectiveness and applicability of our model as evident in the outstanding results of several performance evaluation metrics.

The rest of this paper is organised as follows: Section 2 reviews the theoretical background and existing work related to tweet mining. The framework of the proposed approach is described in Section 3. Section 4 presents the various machine learning algorithms which are incorporated into the proposed framework. The detailed experiments conducted to classify Twitter users and their tweets are described in Section 5. In Section 6, the motivation for the research and the benchmark results are discussed with the state-of-the-art approaches. Finally, the paper is concluded by listing the contributions, the limitations and the anticipated enhancements of the proposed framework.

2. Theoretical background

Since the uprising of Web 2.0, the role of Web browsers has changed to enable users to send and receive content that is leveraged by several online tools such as e-mail applications and chat forums to more recent and revolutionary electronic platforms such as OSNs. OSNs such as Facebook, Twitter, LiveBoon, Orkut, Pinterest, Vine, Tumblr, Google Plus and Instagram among others allow users to share videos, photos and files, and have instant conversations. These platforms provide important means of growing and adhering between societies, bringing together concepts and visions, in addition to its active and distinctive role as an effective medium of social interaction. The dramatic increase in the impact of social data is a testimony to our growing digital lifestyles. Social data have emerged in industries and activities ranging from marketing and advertising to intelligence gathering and political influence. In fact, the extent of this revolution is continually spreading; it is about building data infrastructures that are needed to effectively digest the breeding of social data to achieve added value. This has motivated research communities to dig deep, to provide solutions and to develop platforms for potential use of these data sets in several applications (e.g. marketing [Bolotaeva and Cata, 2010], e-commerce [Kaplan and Haenlein, 2010], education [Tess, 2013], health [Salathé *et al.*, 2013], etc.]. These endeavours include the recent efforts to understand the dynamic and unstructured nature of social content in an attempt to deliver the right content to its interested users (De Maio *et al.*, 2017; De Maio *et al.*, 2017). Furthermore, social media has also been used to improve employees' productivity (Scuotto *et al.*, 2017), knowledge sharing (Bolotaeva and Cata, 2010; Scuotto *et al.*, 2017; Ardichvili *et al.*, 2006) and overall firm innovation performance (Scuotto *et al.*, 2017).

The following section discusses the theoretical background for the current approaches to Twitter mining followed by an evaluation to these approaches and the proposed solutions.

2.1 Semantic data analysis

Berners-Lee introduced the notion of the semantic Web to facilitate the machine understanding of Web language; these data can be used across several applications (Berners-Lee *et al.*, 2001). Ontology is defined as the formal explicit specification of a shared conceptualisation (Gruber, 1995). Incorporating semantic analysis in the area of social big data has generated a steady support from several research communities. Such endeavours attempt to untangle the ambiguity of the unstructured nature of social data

content and discover the domain of knowledge through incorporating semantic analysis techniques to identify, annotate and enrich entities embodied in social data content. In other words, incorporating semantic analysis ushers a better understanding of the contextual content of social media data through extracting their semantic data. De Nart *et al.* (2015) proposed a content-based approach to extract the main topics from the tweets. This approach is an attempt to understand the research communities' activities and their emerging trends. Chianese *et al.* (2016) proposed a data-driven and ontology-based approach to identify cultural heritage key performance indicators as expressed by social network users. This approach can be used in different domains but is only relevant to user domains. Iwanaga *et al.* (2011) and Ghahremanlou *et al.* (2014) both applied ontology to create applications in crisis situations. The former ontology was designed to be used for earthquake evacuation to help people locate evacuation centres based on data posted on Twitter. The latter showed a geo-tagger that aims to process unstructured content and infer locations with the help of existing ontologies. Kumar and Joshi (2017) harnessed the ontology-driven approach to obtain Twitter users' interests; however, their experiments were conducted at the tweet level only, lacking a consideration of the user's domain of interest. Twitter mining through semantic analysis has been further extended to address social media trends (Kaushik *et al.*, 2015), sentiment analysis (Saif *et al.*, 2016), knowledge base and discovery (Sendi *et al.*, 2017), employment trends (Mehta and Buch, 2016), event classification (Romero and Becker, 2017) and fundamentalism detection (Saif *et al.*, 2017), among others.

2.2 Machine learning for data classification and topic distillation

Topic distillation (a.k.a topic discovery, topic modelling, latent topic modelling or statistical topic modelling) is an automatic approach used to distil topics from a corpus of words embodied in a set of documents incorporating statistical techniques (Blei *et al.*, 2003; Anthes, 2010; Wang *et al.*, 2009). The primary reason for developing topic discovery techniques is to improve information retrieval particularly when searching large corpora of data and indexing.

These statistical-based techniques have also been used as other means of topic modelling and discovery in social data mining. Examples of such statistical-based techniques are LDA (Blei *et al.*, 2003), latent semantic analysis (LSA), and more recently, fuzzy latent semantic analysis (FLSA) (Karami *et al.*, 2017). LDA is based on an unsupervised learning model harnessed to identify topics from the distribution of words. LSA, an early topic modelling method, has also been extended to pLSA (Hofmann, 1999), and generates the semantic relationships based on a word-document co-occurrence matrix. FLSA supposes that the list of documents and their embodied words can be fuzzy clustered, where each cluster is represented by a certain topic. LDA and similar unsupervised techniques have been widely used in several modelling applications (Chen *et al.*, 2016; Nichols, 2014; Weng *et al.*, 2010; Asharaf and Alessandro, 2015; Quercia *et al.*, 2012; Onan *et al.*, 2016). Vicient and Moreno (2015) presented a methodology for unsupervised topic discovery through linking social media hashtags to terms of WordNet. Furthermore, Alam *et al.* (2017) harnessed in their approach statistical techniques that are able to detect interpretable topics. Incorporating statistical techniques to benefit social data analysis approaches is also evident in the literature; Twitterrank (Weng *et al.*, 2010) incorporates the LDA technique to classify users' interests through applying the LDA modelling technique to the overall content of each user. Ito *et al.* (2015) adopted LDA for topic discovery to validate the credibility of the content on Twitter. Xiao *et al.* (2013) proposed an approach for predicting users influence in the social data context. They computed the topic distribution of users through the use of LDA technique.

2.3 Evaluation of current approaches

2.3.1 Inclusion of both user and tweet levels. The increasing use of Twitter has motivated researchers to develop several methods for discovering the main interest (s) of their users. Because of the ambiguity, shortness and nosiness of tweets (Michelson and Macskassy, 2010), these endeavours are still in their infancy; hence, extensive research in this area is vital (Shen *et al.*, 2015). Twitter tools (Sherchan *et al.*, 2013; Saif *et al.*, 2016; Russell, 2003) are focused on the exploration of user networks to obtain information for user interests and topics. These approaches only extract keywords to obtain a summary of Twitter data. However, the use of keywords only cannot fully cover user domains and may generate misleading user information. Therefore, the proposed approach in this study considers both the user level and tweet level, which involves semantics of words and accurate disambiguation for social networks study. The accurate classification of the users' interest assists in providing an accurate understanding of short textual content of future tweets. This benefits several applications, the aim of which is to obtain a correct domain-based trustworthiness of users and their content in OSNs.

2.3.2 Integration of different repositories. There have been two main research avenues in which domains of interest have been investigated and inferred from the textual content of users in OSNs. The first avenue focuses on the incorporation of ontologies, semantic Web and linked data to enrich textual data and extract knowledge, thereby linking the textual data with a particular user domain. For instance, Michelson and Macskassy (2010) used the DBpedia knowledge base to annotate entities in users' tweets, and extract the users' main interests by using the categories proposed on Wikipedia. De Maio *et al.* (2017) used Wikipedia to infer users' topics of interest embodied onto their proposed ranking algorithm. Wikipedia has also been used as a knowledge base repository for topic discovery in Schonhofen's (2006) and Hassan *et al.*'s study (2012). In addition to DBpedia, the current approach incorporates other knowledge base repositories, such as Freebase, YAGO and OpenCyc. Furthermore, the study adopts and extends the BBC Politics ontology to capture politics domain knowledge.

2.3.3 Incorporation of domain ontology, semantic Web and machine learning. Statistical techniques have been used as another means of topic modelling and discovery in Twitter mining. The two dominant statistical techniques that have been used are LDA (Blei *et al.*, 2003) and LSA. LSA, an early topic modelling method that has been extended to pLSA (Hofmann, 1999), generates the semantic relationships based on a word-document co-occurrence matrix. LDA, is an extension of pLSI, and LDA is based on an unsupervised learning model to identify topics from the distribution of words. These approaches have been widely used in several modelling applications (Chen *et al.*, 2016; Nichols, 2014; Weng *et al.*, 2010; Asharaf and Alessandro, 2015). However, the high-level topic classifications that use these bag-of-words statistical techniques are inadequate and inferior (Michelson and Macskassy, 2010). Furthermore, the brevity and ambiguity of such short texts make the process of topic modelling using these statistical models more difficult (Li *et al.*, 2016). In addition, these methods do not consider the temporal factor. In other words, the users' knowledge evolves over time and their interest might be diverted elsewhere depending on their experience, work, study or other factors. Hence, it is important to scrutinize users' interest over time to infer intrinsic topics of interest to users in OSNs. The approach of this study addresses these problems through the use of a systematic process which addresses temporally the domain of interest at the user level, and attempts to identify the domain not readily evident at the tweet level. The approach includes the use of domain ontology, semantic Web technologies and machine learning, where domain ontology and semantic Web attempt to extract the semantics of textual data to determine the domain of the textual data, and machine learning attempts to perform domain-based classification at the user and tweet levels.

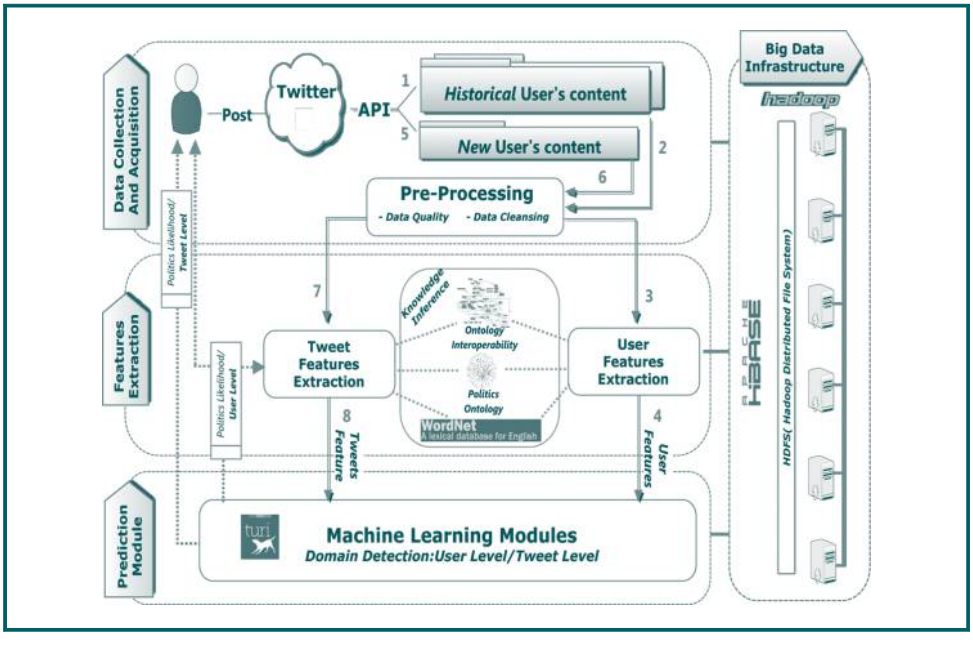
2.3.4 Addressing key features of big data. There is a notable consensus among researchers that the traditional tools for collecting, analysing and storing data are no longer able to handle large amounts of big data. Therefore, more advanced, unconventional and adaptable technical solutions are required to address the challenges of managing a wide variety of big data islands, which expand exponentially through the huge data generated from tracking sensors, OSNs, transaction records, metadata and many other data fountains. Manyika *et al.* (2011) listed some of the big data technologies, such as Big Table, Cassandra (open Source DBMS), Cloud Computing and Hadoop (open source framework for processing large sets of data). Chen *et al.* (2014) discussed various open issues and challenges of big data, and listed the key technologies of big data. The incorporation of big data technology to facilitate domain discovery and to measure users' trustworthiness in OSNs is unavoidable, particularly regarding the nature of the content of social media which has a wide berth. This has interestingly attracted researchers of social trust to leverage the big data techniques to benefit their conducted experiments (Lavbič *et al.*, 2012/2013; Herzig *et al.*, 2014; Smith *et al.*, 2012). However, previous studies have failed to address the key features of big data, such as volume (i.e. massive social data data sets), veracity (i.e. reputation of the data sources) and value (outcome product of the data analytics). Hence, this study starts with the characteristics of big data and sorts out issues related to these dimensions to better obtain the anticipated outcomes of social data analysis.

3. System architecture

Figure 1 shows the architecture of the proposed framework which adopts a big data infrastructure. This framework comprises three main components, namely, data collection and acquisition, features extraction and the prediction module. The big data infrastructure at the School of Information Systems, Curtin University, is used as a distributed environment to facilitate data storage and analysis. This facility has a six-node cluster, each with 2 TB storage, 8 core processors and 64 GB RAM.

The information flow through the proposed framework can be described in steps. As shown in Figure 1, Steps 1-4 represent the processes required to achieve the predicted likelihood

Figure 1 System architecture



value of the user's interest in the politics domain. This is the first outcome value (politics likelihood/user level) indicated by the red-dotted line. Steps 5-9 follow and predict the politics domain-based likelihood value of a newly posted user tweet. This is the second outcome value (politics likelihood/tweet level) indicated by the red-dotted line. In the proposed framework, the user posts public content to the Twitter network, which facilitates data collection through the available application programming interfaces (APIs). The user's content is collected in two phases, namely, historical user's content and new user's content. The user's historical content represents the recent and former tweets which are collected in the first phase. The user's new content refers to their future tweets which will be collected during the second phase.

The collected historical tweets are pre-processed and passed to either the tweet features or user features extraction module. A list of user features is extracted and fed into a machine learning module to predict the politics domain likelihood value, where the domain likelihood indicates the user's interest in the politics domain. This domain likelihood is harnessed further and is added as another feature to the list of features extracted from the new user tweet after pre-processing during the second phase. The newly combined list of tweet features is fed into the machine learning module to predict the politics domain likelihood of the newly posted tweet. The following subsections explain the mechanism of each component of the proposed framework.

3.1 Data collection and acquisition

3.1.1 Data generation and selection. Since the establishment of Twitter™ in 2006, Twitter has provided a rich data set of over 500 million tweets daily which is around 200 billion tweets a year (Sayce, 2016). Twitter mining is an emerging research field falling under the umbrella of data mining and machine learning. Twitter™ is the chosen subject of this paper because of the following reasons:

- Twitter is a fertile medium for researchers in diverse disciplines, leveraging the vast volume of content.
- Twitter facilitates data collection by providing easy access APIs to the Twitter sphere.
- It is challenging to determine the accurate domain (s) to which the user's tweet is referring because of the economy and the ambiguity and brevity of a tweet's content,.

For the purpose of proof of concept, this study is limited to an on/off domain classification to the content of OSNs. Hence, the politics domain has been selected for the following reasons:

- Twitter has been intensively incorporated as an important arena by politicians to express and defend their policies, to practice electoral propaganda and to communicate with their supporters (Shapiro and Hemphill, 2017).
- Twitter has raised considerable controversy regarding its usage as a platform to attack political opponents (Van Kessel and Castelein, 2016).
- Twitter is characterised by its growing social base to include broad political social groups leveraged by ease of use, free access and deregulated nature (Halberstam and Knight, 2016).
- The amount of the political discourse in social content is overwhelming; over one-third of OSNs' users believe that they are worn-out by the quantity of the political content they encounter (Duggan, 2016).

Such an abundance of data facilitates data aggregation and improves the outcome of data analysis. For future work, this study aims to develop a multi-domain-based classification,

leveraged by domain ontologies, semantic technologies and linked open data. Hence, beside the politics domain, an analysis of other domains of interest may be further investigated in the future.

The data set used for this study has been collected using Twitter's "User_timeline[3]"API method. This mechanism allows access to and retrieval of public users' content and metadata. The collection of the users' content was accomplished in two stages:

1. By collecting historical user content (up to "3,200" most recent tweets[4]). This data set will be used to predict the user's interest in the politics domain in general.
2. By collecting the new content of those users whose historical tweets were obtained in the first phase.

This is used to predict the politics domain likelihood value of the new tweet. As will be described later, the data set of the first stage is used to predict the user's interest in politics at the user level, i.e. to establish an understanding of the user's interest in the politics domain based on the user's past content. The politics domain likelihood value of the new user's tweet is predicted on the basis of the analysis of its content, other than the politics interest likelihood value predicted at the user level.

3.1.2 Pre-processing data. The veracity of data refers to the certainty, faultlessness and truthfulness of data (Demchenko *et al.*, 2013). Although reliability, availably and security of data's nascence and storage are significant, these factors do not guarantee data correctness and consistency. Appropriate data cleansing and integration techniques should be incorporated to ensure the certainty of data. The data collected for the user's content, and historical and new tweets, are pre-processed by data quality enhancement and data cleansing techniques which are discussed below:

- *Data cleansing* of user content is conducted by using the following techniques: all redundant content (i.e. same data set crawled more than once) such as tweets or user data is eliminated with their metadata; removing stop words; removing URLs; decoding all HTML entities to their applicable characters; eliminating all HTML tags such as <p>, <a>, etc.; removing punctuation marks; correcting encoding format, etc.
- *In data quality enhancement*, the list of Twitter handles (a.k.a. Twitter user/screen name such as @example), which are indicated in the user's tweets, is collected and replaced with the user's corresponding names. This is achieved through the API of Twitter's "lookup[5]". These handles are normally neglected or deleted when mining user's tweets. However, these handles are important because these are used by Twitter users to mention other Twitter users in their tweets, replies or re-tweets. Hence, it is essential to identify and ascertain the actual names of those users. This assists in the process of domain extraction. For example, a user shows an interest in the politics domain if she/he commonly indicates handles linked to politicians or political parties, in addition to publishing other politics-related content.

3.2 Features extraction

The pre-processed data set is passed to the features extraction modules. For the new users, the features of their content (historical tweets) are extracted in the "User Features Extraction" module. As for the new tweets of the already existing users, features are extracted in the "Tweet Features Extraction" module.

The aim of this study is to establish a fundamental ground for efficiently detecting the domain of interest of Twitter users, which will significantly contribute to a better understanding of the domain (s) of future users' tweets. As a proof of concept, the proposed system is validated by an application on the politics domain, where the proposed system attempts to detect whether the domain of a tweet is or is not politics-related. This

validation is based primarily on former knowledge about a user's political interests obtained by analysing the user's historical content. To do so, the following politics-domain knowledge inference approach is designed to extract the semantics of a user's tweets, thereby uncovering the user's domain of interest.

3.2.1 Politics domain knowledge inference. In the feature extraction module, domain knowledge inference is the main process used to extract user and tweet features from pre-processed data sets. For the purpose of proof of concept, the study focuses on the politics domain, using politics ontology, WordNet and ontology interoperability to infer politics knowledge.

3.2.1.1 Politics ontology and WordNet®. The politics domain refers to the knowledge captured in politics ontology along with its knowledge base. BBC defines politics ontology as "an ontology which describes a model for politics, specifically in terms of local government and elections" (BBC, 2014). The BBC Politics ontology conceptualises a politics model especially for the UK Government and elections. It was originally designed to cope with UK local government and European Elections in May 2014. This study applies the BBC Politics ontology to Australian politics by further extending politics concepts. Figure 2 shows the BBC Politics ontology, whereas Figure 3 shows the Politics ontology used in this research. Furthermore, the study uses WordNet[6], which is a lexical dictionary used to construct relations between terms through synonymies. Synonyms (or synsets) are a set of interrelated terms or phrases which indicate the same semantic concept, such as the words "elections, public opinion poll, opinion poll and ballot". All the synsets of the political concepts captured in politics ontology depicted in Figure 3 are examined, and only the synonyms applicable to the politics context are captured.

3.2.1.2 Ontology interoperability. The interlinking with other relevant entities defined in other data sets supports interoperability. The approach taken in this study addresses information interoperability by focusing on the equivalence links that direct the URI to refer to the same resource or entity. The politics ontology supports the equivalence links between the ontology components and the tweet data. The resources and entities are linked through the owl#sameAs relation. This implies that the subject URI and object URI resources are the same, and hence the data can be further explored.

In the interlinking process, we incorporate AlchemyAPI™ as a one-stop shop, leveraging access to a wide variety of linked data resources[7] through providing easy access to APIs. These resources include but are not limited to different vocabularies, such as Upper Mapping and Binding Exchange Layer, Freebase (which is a community-curated database for well-known people, places and things), YAGO high-quality knowledge base and DBPedia knowledge base. These resources are used to help extend the knowledge base of the politics ontology by identifying (non-)Australian politicians and (non-)Australian political

Figure 2 BBC Politics ontology

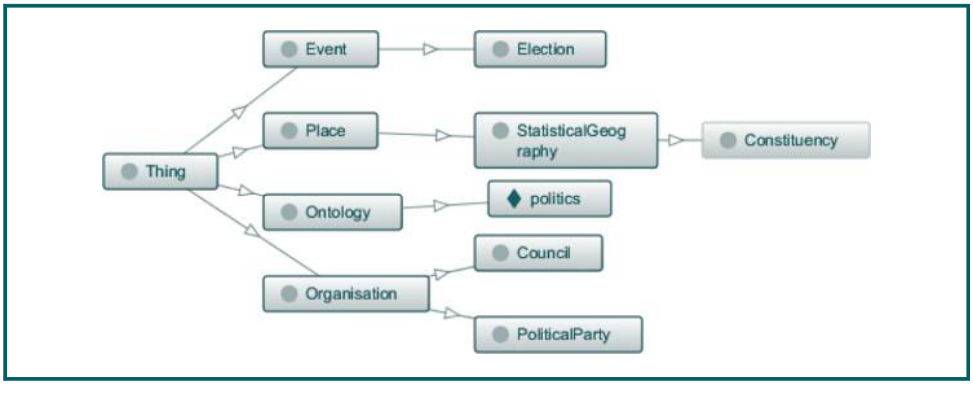
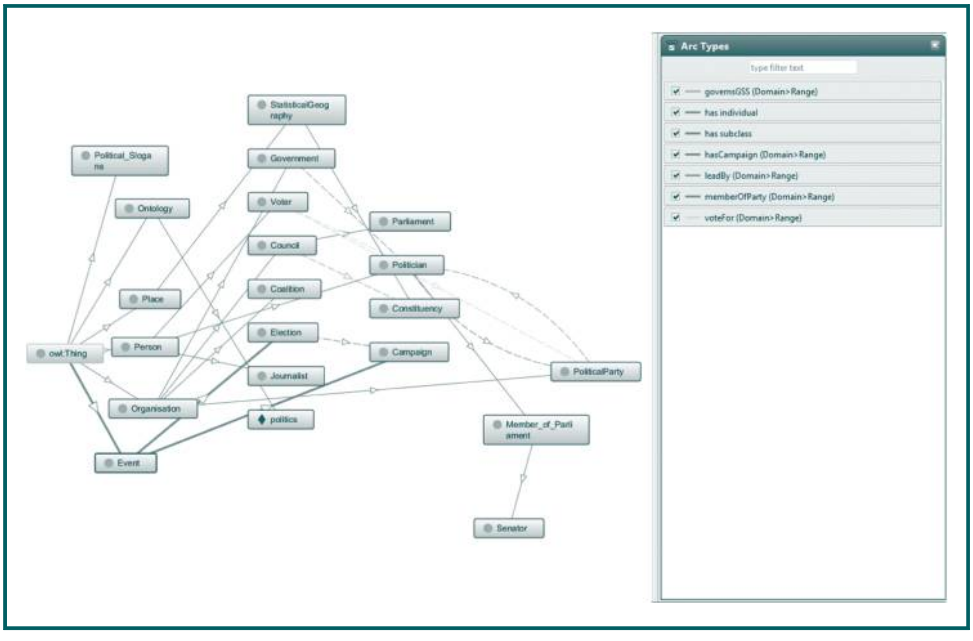


Figure 3 BBC Politics ontology extension



parties from users' tweets. For example, at this stage, we capture "99,812" instances of "2009" politicians, and "48,704" instances of "59" political parties in the politics ontology.

3.2.2 User-level features. The political interest of users is primarily measured by two main proposed factors: continuity and knowledgeability. Continuity refers to the frequent interest of a user in a certain domain. In other words, the user demonstrates an interest in the politics domain by tweeting or retweeting content in this domain over a relatively long period of time. Continuity is measured by counting the number of political entities identified from the user's tweets in each time period (such as every month, quarter, etc.). Knowledgeability (or speciality) refers to the user's close acquaintance with the politics domain and also refers to the user's dedicated pursuit of the politics domain through a commitment, such as work or study. Knowledgeability is measured by accumulating the distinct number of political entities annotated from the user's tweet, and the user's profile description. Table I shows the list of features used to classify whether the user's interest is *on-topic* or *off-topic*.

Table I A list of user's features

No.	*Features*	*Description*
1	no_tweets, x_1	The total count of users' historical collected tweets up to 3,200 tweets
2	unq_pol_entities, x_2	Total count of distinct/unique political entities extracted from all user's tweets
2	pol_entities_pre_QW_YYYY, x_3	Count of political entities annotated from the tweets posted before quarter "W" of the year "YYYY"'
3	pol_entities_QW_YYYY, x_4	Count of political entities annotated from the tweets posted in quarter "W" of the year "YYYY"
4	pol_entities_QX_YYYY, x_5	Count of political entities annotated from the tweets posted in quarter "X" of the year "YYYY"
5	pol_entities_QY_YYYY, x_6	Count of political entities annotated from the tweets posted in quarter "Y" of the year "YYYY"
6	pol_entities_QZ_YYYY, x_7	Count of political entities annotated from the tweets posted in quarter "Z" of the year "YYYY"
7	profile_pol_entities, x_8	Count of political entities annotated from user's profile description
9	verified (Authentication Status), x_9	Authentication flag used for accounts of public interest (for example, politicians)

On-topic refers to when the user demonstrates a continuous interest in the politics domain. *Off-topic* users are those whose Twitter content shows their non-interest in the politics domain.

Features x_2 to x_8 as depicted in Table I are selected to primarily focus on users' ongoing interest in and knowledge about the politics domain by extracting the political entities from their tweets and by leveraging the knowledge-inference tools explained in the previous section. In particular, features x_2 to x_8 are proposed to address the political knowledgeability of users. Moreover, features x_3 to x_7 address the continuing interest of users in the politics domain. Features x_1 and x_9 are added to support the aforementioned features and will be discussed later in this paper.

Unq_pol_entities (x_2), listed in Table I, refers to the number of distinct political entities extracted from the history of a user's tweets. Profile_pol_entities (x_8) represents the number of all political concepts that are identified in the users' profile description on their Twitter accounts. The former feature represents the diversity of the political concepts embodied in the users' tweets, and the latter feature, x_8 is used to examine the explicit indication of the users' interest in the politics domain, particularly if the users work in this domain. This is usually clearly indicated in their profile description.

The list of all political entities is counted periodically. The political entities extracted from the user content for each time period is used to scrutinise political interest temporally rather than scrutinising the tweets as a whole. Therefore, the collected historical tweets are divided into five groups: x_3 to x_7. Four groups, x_4 to x_7, indicate the four sequential and recent quarters (W, X, Y and Z), where "Z" is the most recent quarter, and one group, x_3, indicates the rest of the tweets posted before the "W" quarter. This mechanism is proposed because the users' interest (s) may change, and their knowledge may evolve over time. Hence, it is more efficient to examine the user's domain (s) of interest based on current and recent behaviours from the four time groups. Furthermore, some users only show a particular interest in the politics domain when popular political events are taking place. For example, a user's involvement in conversations during election campaigns does not necessarily indicate an interest in the politics domain generally, as the election is a trending topic only, on which users with dissimilar interests share their thoughts, and/or anticipations about the potential candidates.

The remaining two features listed in Table I are the no_tweets, and verified features. The no_tweets, x_1, relates to the number of collected historical tweets. This feature is important as a means of addressing the ratio between the number of political concepts accumulated for Features x_2 to x_8 and the total number of tweets. For example, two users might archive the same number of distinct political concepts, although the number of tweets differs for each user. The verified feature, x_9, is the authenticated flag (i.e. blue verified badge). Twitter may set this flag to "1" for users of public interest. Twitter currently offers this feature to help users find influential and high-quality accounts in several domains.[8]

3.2.3 Tweet-level features. In the previous section, the user's historical collected tweets were studied to obtain an accurate understanding of that user's interest in the politics domain. A list of features extracted from the content at the user level is formulated and will be used to predict the user's political interest (likelihood). On this backdrop, the likelihood of the user's interest in the politics domain would be a main driver facilitating an understanding of the domain of the users' future tweets. Table II summarises the list of features selected to predict the political likelihood at the tweet level.

As shown in Table II, political_entities (x_{10}) represents the number of political entities annotated from the tweet using the aforementioned knowledge discovery tools. Words_count (x_{11}) is the number of remaining words in the tweet after the cleansing process. Political_perc (x_{12}) represents the ratio between the number of political entities annotated in the tweet to the total words used. Despite its brevity, a tweet might discuss

Table II A list of tweet features

No.	*Features*	*Description*
1	political_entities, x_{10}	Count of political entities extracted from the tweet
2	words_count, x_{11}	Count of tweet's words
2	political_perc, x_{12}	Computed as $\frac{x_{10}}{x_{11}}$
3	pol_entities_recent_quarter, x_{13}	Count of political entities annotated from the user's tweets posted in the most recent quarter
4	user_pol_likelihood, x_{14}	Political likelihood value

more than one topic; thus, x_{12} is proposed as an indicator of the weight of the politics domain in the tweet. Pol_entities_recent_quarter (x_{13}) represents the number of political entities from all tweets posted during the most recent quarter. This feature is included because it represents the user's most recent political (non-)interest. User_pol_likelihood (x_{14}) is the predicted value obtained from user analysis which signifies a user's general interest in the politics domain.

Features x_{13} and x_{14} are proposed to indicate the recent political interest of the user. These features assist in further understanding the actual context of the newly posted tweet, given their typically short length and ambiguity. Hence, users who have been predicted to be interested in the politics domain will likely post politics-related content in future posts. This will be discussed and demonstrated further in the experiment section (Section 5).

4. Machine learning module for classification

This section provides an overview of well-known machine learning classification algorithms. Based on the user and tweet features, $\overline{x} = [x_1, x_2, \quad \text{to } x_{14}]$, a machine learning module determines the likelihood of whether or not a user/tweet is in the politics domain, namely, *y*, where the following commonly used implicit or explicit classifiers including logistic regression (LR) (Hosmer *et al.*, 2013), decision tree (Quinlan, 1993; Ho, 1995; Friedman, 2001) and support vector machine (SVM) (Boser *et al.*, 1992) are used for user-based classification, and LR is used for tweet-based classification. For demonstration purposes, this overview will consider the domain-based classification at the user level. LR (Al-Tahrawi, 2015; Yen *et al.*, 2011), decision tree (Sharef *et al.*, 2015) and SVM (Altınel *et al.*, 2015; Dong *et al.*, 2016) in particular have been used for text categorisations. Also these approaches are more narrow and computationally simpler than recently developed machine learning approaches, such as the deep learning or deep networks approaches.

Development of a novel classifier is not the main research focus of this paper. Hence, the study attempts to implement a computationally simple but effective approach. Five commonly used approaches are used, namely, LR, SVM, top-down inducing based decision tree (TD-DT), random forest-based decision tree (RF-DT) and gradient-boosting-based decision tree (GB-DT).

4.1 Logistic classifier

Logistic regression is commonly used for conducting binary classification tasks (Hosmer *et al.*, 2013). In LR, the likelihood of whether the user is in the politics domain is determined by a logistic function consisting of a linear summation of x_1 to x_9. The logistic function is given as:

$$f^{LR}(\overline{x}) = P(y = 1|\overline{x}) = \frac{1}{1 + \exp\left(-\left(b_0 + \sum_{i=1}^{14} b_i \cdot x_i\right)\right)} \quad (1)$$

In the study, b_0, b_1 to b_{14} are the logistic coefficients, which are determined by maximizing the likelihood when y = 1, which indicates that the user is definitely in the politics domain.

Unlike linear regression which has normally distributed residuals, ordinary least square regression cannot be applied to determine the logistic coefficients. Hence, to determine b_0, b_1 to b_{14}, Newton's method is used. Newton's method begins with tentative logistic coefficients and it adjusts the coefficients slightly to see whether these can be improved. It repeats this iterative process until the process converges. A user is classified in the politics domain, when the value of $f^{LR}(\overline{x})$ in (1) is more than 0.5. Otherwise, the user is classified as being in the non-politics domain.

4.2 Support vector machine

Support vector machine is commonly used for conducting binary classification tasks (Boser *et al.*, 1992) particularly involving with the confusion matrix analysis (true-positive [TP] and false-negative [FN]). SVM is relatively new and was designed for applications involving text categorization and recognition (see for example Altınel *et al.*, 2015; Dong *et al.*, 2016).

In SVM, the user is classified as either being in the politics or in the non-politics domains, based on the following formulation:

$$f^{SVM}(\overline{x}) = \text{sgn}(D(\overline{x})) \tag{2}$$

$$\text{where } D(\overline{x}) = \sum_{i=1}^{14} w_i \varphi(x_i) + b; \tag{3}$$

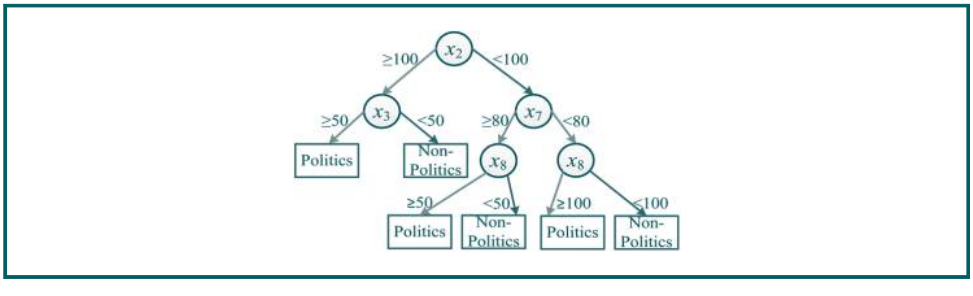
$$\text{and } \text{sgn}(D(\overline{x})) = \begin{cases} 0 \text{ if } D(\overline{x}) < 0 \\ 1 \text{ if } D(\overline{x}) \geq 0 \end{cases} \tag{4}$$

φ is the transform function which is correlated to the kernel function and w_i with $i = 1, 2$ to 14 and b represents the SVM parameters. The five common kernel functions are linear function, homogeneous polynomial, inhomogeneous polynomial, gaussian radial basis function and hyperbolic tangent. The kernel function is generally determined by a trial and error method. After the kernel function has been determined, w_i and b are reformulated as a quadratic programming problem, which is solved by the gradient descent algorithm. When the value of $f^{SVM}(\overline{x})$ in (2) is equal to 1, the user is classified as being in the politics domain. Otherwise, the user is classified as being in the non-politics domain.

4.3 Decision tree classifier

A decision tree is a classifier which can express a recursive partition of the instance space. A decision tree is a flow chart-like structure, where each internal (or non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test and each leaf (or terminal) node holds a class label. The highest node in the tree is the root node. Figure 4 illustrates how a decision tree is used to determine whether a user is in the politics domain.

Figure 4 Example of a decision tree for the politics domain



The study considers a simple decision tree with four features, x_2, x_3, x_7 and x_8. The red branches of the decision tree indicate that the user is in the politics domain; this occurs if any of following three conditions are met:

1. if x_2 is more than 100 and x_3 is more than 50;
2. if x_2 is less than 100, x_7 is more than 80 and x_8 is more than 50; or
3. if x_2 is less than 100, x_7 is less than 80 and x_8 is more than 100.

Compared with LR and SVM, decision trees are very intuitive and easy to interpret. In addition, empirical results have demonstrated that decision trees outperform SVM and LR on 11 benchmark problems, in terms of ten classification metrics (Caruana and Niculescu-Mizil, 2006). Three commonly used approaches, namely, top-down inducing C4.5 (Quinlan, 1993), random forest (Ho, 1995) and gradient boosting (Friedman, 2001), are used to develop decision trees for determining whether a user is in the politics domain. In top-down inducing C4.5, the decision tree is constructed from the top to the bottom, based on a divide-and-conquer mechanism. The top-down inducing C4.5 trains the samples based on the splitting measures. After the selection of an appropriate split, which results in a minimum classification error, each node further subdivides the training samples into smaller subsets of samples, until the split gains satisfy the splitting measure. In a random forest, multiple trees are generated on the basis of randomly selected subspaces of features. The trees generalise their classification in complementary ways and their combined classification attempts to improve each single tree. In gradient boosting, a base decision classifier is expanded by adding additional branches to the base of the tree. The expansion continues until no further improvement can be obtained by adding an additional branch.

5. System evaluation

In Sections 3 and 4, a system framework is proposed to detect the domain-based interest of users/tweets by incorporating machine learning. This section evaluates the effectiveness of the proposed system framework.

5.1 Data sets collection and ground truth

To evaluate the proposed system framework, a list of Australian Twitter users and their public content is collected and pre-processed as discussed in section 3.1. The tentative list of users who are potentially interested in the politics domain is selected from the following sources:

- a list of Members of Parliament and Senators indicated on the official website of the Parliament of Australia [9];
- members and subscribers of three politics-based Australian Twitter lists [10]; and
- miscellaneous sources[11].

Because of the lack of online sources indicating those users interested in politics in OSNs, the aforementioned lists are selected because it is assumed that these people are interested in the politics domain as is evident later in the paper.

Users who are assumed to have little or no interest in the politics domain were tentatively selected from the two collected data sets:

1. members of various Australian Twitter lists established to discuss sports, information technology and other non-politics domains; and
2. a list of Australian users who achieved the highest trustworthiness values in all domains except “news, government and politics”, extracted from an on-going project, the

preliminary approach of which has been described in previous work (Abu-Salih *et al.*, 2015).

The tentative selection criterion is established on the basis of the users' profile description, choosing users who indicated a non-politics interest.

The collected and cleansed tweets of each user is then carefully examined to obtain an accurate understanding of the user's domain of interest, thereby establishing a truth data set for developing and validating the proposed system framework at the user level. In this data set, users are labelled and assigned to two categories:

1. *on-topic* users who show a particular interest in the politics domain; and
2. *off-topic* users who demonstrate no or minimal interest in the politics domain.

Table III shows a tentative list of collected users, and the actual number of users selected for the ground truth, based on an examination of all tweets.

The collected users of the ground truth data set indicated in Table III are analysed with their historical tweets to develop the prediction model. This is used to predict the likelihood of users in the politics domain.

The next phase involves conducting experiments at the user level to predict the politics classification of the new users' tweets. Therefore, another data set is collected which contains new tweets posted by already-examined users. The new tweets are examined and a subset of the tweets is selected to construct the ground truth for conducting experiments at the tweet level. The selection is based on four criteria:

1. tweets indicating a *politics* domain, and posted by *politics* users;
2. tweets indicating a *politics* domain, and posted by *non-politics* users;
3. tweets indicating a *non-politics* domain, and posted by *politics* users; and
4. tweets indicating a *non-politics* domain, and posted by *non-politics* users.

These four criteria are chosen to support the prediction model which will be constructed at the tweet level. Table IV shows the total number of tweets collected on the basis of the four selection criteria.

The proposed system framework is implemented in the Turi Graphlab Create™ which is used for these experiments using the Python programming environment. Turi Graphlab Create is used as it is scalable and can therefore accommodate relatively huge data sets. The proposed system framework is used to conduct the experiment at the user level with

Table III Ground truth – user level

	#Collected users (tentative list)	*Ground truth*
On-topic	310	227
Off-topic	350	283

Table IV Ground truth – tweets level

	Politics users	*Non-politics users*
Politics tweets (on-topic)	150	125
Non-politics tweets (off-topic)	105	100

the nine features (x_1 to x_{14}) illustrated in Table I and the five classifiers discussed in Section 3.3, LR, SVM, TD-DT, RF-DT and GB-DT. Turi Graphlab Create is also used to conduct experiments at the tweet level with the features listed in Table II. Tenfold cross-validation is used on the data sets to evaluate the generalisation capability of the proposed system framework which is embedded with the five classifiers.

At the user-level analysis, and as depicted in Figure 5, the proposed system framework can be used to determine (classify) whether or not a user is interested in politics. The circled ones are classified as the politics-interested users and the non-circled ones are the users who are not interested in politics. Four scenarios are illustrated by the classification as:

1. *true-positives,* which indicate the number of actual politics users that are classified correctly as politics users;
2. *false-positives,* which indicate the number of non-politics users that are classified incorrectly as politics users;
3. *false-positives* (FN), which indicate the actual politics users that are classified incorrectly as non-politics users; and
4. *true-negative* (TN), which indicate the non-politics users that are classified correctly as non-politics users.

These four scenarios can also be shown in the confusion matrix (Table V) which depicts the performance of the prediction. The model illustrated in Figure 5 is also applicable to the tweets classification which is the second analysis phase in the proposed approach.

In Graphlab Create™, the confusion matrix is often a table used to provide further details on the true and false predictions. This table comprises three columns:

1. Target_label: the classification label of the ground truth. It represents the *on-topic* and *off-topic* label in this study;
2. Predicted_label: the classifier prediction label; and

Figure 5 Classification of politics/non-politics users

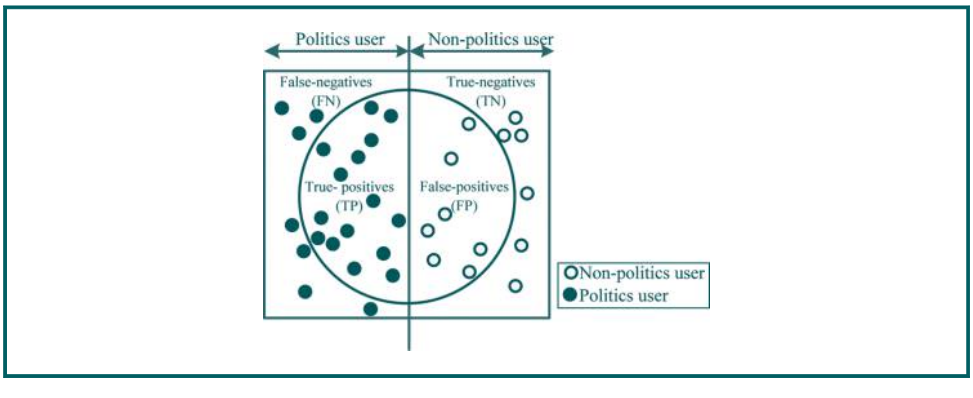


Table V Confusion matrix

	Prediction *On-topic*	*Off-topic*
True		
On-topic	TP	FN
Off-topic	FP	TN

3. Count: the number of times the predicted_label matches the target_label.

The evaluation has been performed by using the following metrics to evaluate the classification performance in predicting whether or not the user/tweet is in the politics domain.

Accuracy indicates the correctness of the incorporated classifier in making the correct prediction. This is essentially the ratio between the correct predictions (i.e. TP + TN) and the total predictions (FN + TP + FP + TN). This is computed as:

$$Accuracy = \frac{TP + TN}{FN + TP + FP + TN} \tag{5}$$

Log-loss (logarithmic loss) is a fine-grained classification evaluation metric. This value is computed by the negative of the accumulation of the log probability of each sample, normalised by the number of samples:

$$Log - Loss = -\frac{1}{n}\sum_{i \epsilon 1,..N} \left(y_i \log(P_i) + (1 - y_i)\log(1 - P_i)\right) \tag{6}$$

Where y_i is the i-th target value, and P_i is the i-th predicted probability. This metric is used because the likelihood probability is addressed to predict the *on-topic* or *off-topic* likelihood of the user or tweet.

Precision, Recall and F-score are metrics commonly used to evaluate classification performance. Precision, Recall and F-score are shown in equations (7), (8) and (9), respectively.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{9}$$

Precision indicates the ratio between the number of actual politics users/tweets that are classified correctly, and the total number of correct and incorrect classifications of politics users/tweets. Recall indicates the ratio between the number of actual politics users/tweets that are classified correctly, and the total number of actual politics users/tweets. Hence, high precision indicates that the classifier is capable of generating substantially more relevant predictions for actual politics users/tweets than the irrelevant ones. High recall indicates that the classifier is capable of generating most of the relevant predictions for actual politics users/tweets. Precision with a value of "1" indicates that every prediction is the actual politics user/tweet but it does not mean that all the actual politics users/tweets are retrieved, whereas a recall score with a value of "1" indicates that all predictions are actual politics users/tweets but it does not indicate the number of non-politics predictions that are retrieved. Hence, the F-score is used to provide the trade-off between precision and recall.

5.2 Domain detection – user level

The aforementioned features in Table I are analysed for each user where tweets are divided temporally into five groups to address the temporal dimension. The second and third columns in Table VI show the feature values with respect to the on-topic samples and off-topic samples, respectively, where the *on-topic samples* represent the list of users

Table VI Data set statistics – user level

	On-topic samples	*Off-topic samples*	*ARD*
Total #sers	227	283	
Total #Tweets, x_1	499,475	611,014	10.044
Total #unq_pol_entities, x_2	14,818	2,833	67.9
Total #pol_entities_pre_Q3_2015, x_3	110,128	8,770	85.248
Total #pol_entities_Q3_2015, x_4	18,492	869	91.023
Total #pol_entities_Q4_2015, x_5	14,842	522	93.205
Total #pol_entities_Q1_2016, x_6	21,562	601	94.577
Total #pol_entities_Q2_2016, x_7	39,712	1,218	94.048
Total #profile_pol_entities, x_8	*237*	*0*	*100*
Total #Verified, x_9	167	94	27.969

interested in the politics domain and the *off-topic samples* show the users who did not have an interest in the politics domain. For the on-topic samples, the *i*-th feature is denoted as x_i on-topic. For the off-topic samples, the *i*-th feature is denoted as x_i off-topic. Absolution relative difference (ARD) in equation (10) is used to indicate the relative difference between the *on-topic samples* and the *off-topic samples*.

$$\mathrm{ARD} = 100 \times \mathrm{Abs}\left(\frac{x_i^{\text{on\_topic}} - x_i^{\text{off\_topic}}}{x_i^{\text{on\_topic}} + x_i^{\text{off\_topic}}}\right) \tag{10}$$

The higher the ARD value, the higher the impact of the corresponding feature used to discriminate the *on-topic* and *the off-topic* users. For example, an ARD of x_8 equal to 100 indicates that x_8 is highly significant in identifying the (non-)interested users in the politics domain by examining their profile description. This evidence will be discussed later.

As depicted in Table VI, the political entities detected in Features x_2 to x_8 for *on-topic* users are much greater than the entities detected for the *off-topic* users. This is because the *on-topic* users have shown an extensive interest in the politics domain through their content on Twitter.

To evaluate the effectiveness of the proposed system framework embedded with the five classifiers (LR, SVM, TD-DT, RF-DT and GB-DT), tenfold cross-validations are used. In the cross-validations, the total observations (i.e. 510 users) are randomly split into two data sets, namely, the training data set (which is 80 per cent of the total sample) and the validation data set (which is 20 per cent of the total sample). Table VII illustrates the main

Table VII Classifiers settings

Classifier	*Main settings*	*Parameters*
LR	Hyperparameters – L1 penalty	0
	Hyperparameters – L2 penalty	0.01
	Solver	Newton–Raphson
	Solver iterations	9
SVM	Solver	L-BFGS[a]
	Predefined number of iterations	10
	Hyperparameters mis-classification penalty	1
TD-DT	Number of trees	1
	Max tree depth	6
RF-DT	Number of trees	10
	Max tree depth	6
GB-DT	Number of trees	10
	Max tree depth	6

Notes: [a]L-BFGS: It is a limited memory of Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm. This is a robust solver for data sets with many coefficients

settings and parameters used to train each of the five classifiers in the proposed system framework.

Table VIII depicts the confusion table used to quantify the performance of each classifier. It can be seen that the LR performs better in the classification task of this study; of the 107 samples used to validate each algorithm, only two samples were incorrectly classified by LR. However, all other classifiers, TD-DT and RF-DT algorithms for example, wrongly classified more samples in the prediction validations. Nevertheless, the classification performance of the incorporated algorithms is acceptable. These methods can generally perform effectively in terms of this domain classification problem.

Table IX shows the evaluation performance metrics of each classifier, where the means and variances for the tenfold cross-validations are given. The metric means that the non-bracketed values and the metric variances are the bracketed values. It can be seen from Table IX that LR achieves better metric means for the five classification metrics among all of the methods where Accuracy, Precision, Recall and F1_score are "the larger-the-better" and Log_loss is "the smaller-the-better". The metric variances generated by LR are generally the smallest. Therefore, LR can yield the best and most robust classification when compared to the other four methods.

Despite the classifier's convergence on the four metrics (i.e. Accuracy, Precision, Recall and F1-score), LR is generally better than the other four methods, particularly regarding log_loss. This indicates that the predicted likelihoods of the validation data set using FR closely match with the assigned labels. TD-DT on the other hand is generally the poorest method when compared with the others.

Table X shows the highest estimated coefficient values calculated for each feature using LR. It shows that "profile_pol_entities, x_8" is the highest estimated coefficient. This is consistent with the results illustrated in Table VIII, where x_8 has the highest impact when compared with the other features. This is because of the importance of this feature in distinguishing the user's interest in the politics domain. In particular, users whose profile descriptions include politics-related entities, such as a parliament member and political journalist, are likely to suggest the politics domain in their tweets.

Table VIII Confusion table

Target_ label	*Predicted_ label*	*LR*	*SVM*	*TD-DT*	*RF-DT*	*GB-DT*
On-topic	On-topic	*59*	58	41	57	48
Off-topic	Off-topic	*46*	46	52	45	45
On-topic	Off-topic	*2*	3	2	4	3
Off-topic	On-topic	*0*	0	3	1	1

Table IX Performance comparison of five classifiers to detect user political interest

	Accuracy	*Log_loss*	*Precision*	*Recall*	*F1_score*
LR	0.9824 (0.0002653)	0.0406	1.0000	0.9672	0.9833
SVM	0.9784 (0.003417916)	0.5781	1.0000	0.9508	0.9748
TD-DT	0.9157 (0.033453)	0.4816	0.9318	0.9535	0.9425
GB-DT	0.9255 (0.032357)	0.1321	0.9831	0.9508	0.9667
RF-DT	0.9490 (0.009473)	0.2321	0.9828	0.9344	0.9580

Notes: Accuracy, Precision, Recall and F1_score are "the larger-the-better". Log_loss is "the smaller-the-better"

Table X Highest positive coefficients – user level

Feature	*Value*
profile_pol_entities, x_8	8.601
verified, x_9	2.162
unq_pol_entities, x_2	0.144
pol_entities_Q4_2015, x_5	0.02

In addition, the *t*-test (Box *et al.*, 2005) was used to evaluate the significance of the hypothesis that the accuracy means obtained by the best method LR are higher than those obtained by the other methods (SVM, TD-DT, RF-DT and GB-DT). The *t*-values between LR and the other methods are shown in Table XI. Based on the *t*-distribution table, if the *t*-value is higher than 1.699, the significance is 95 per cent confidence, which means that the accuracies obtained by the LR are higher than those obtained by the other methods with a 95 per cent confidence level. The *t*-value can be determined by:

$$t\text{–value} = \frac{\mu_2 - \mu_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)/N}} \tag{11}$$

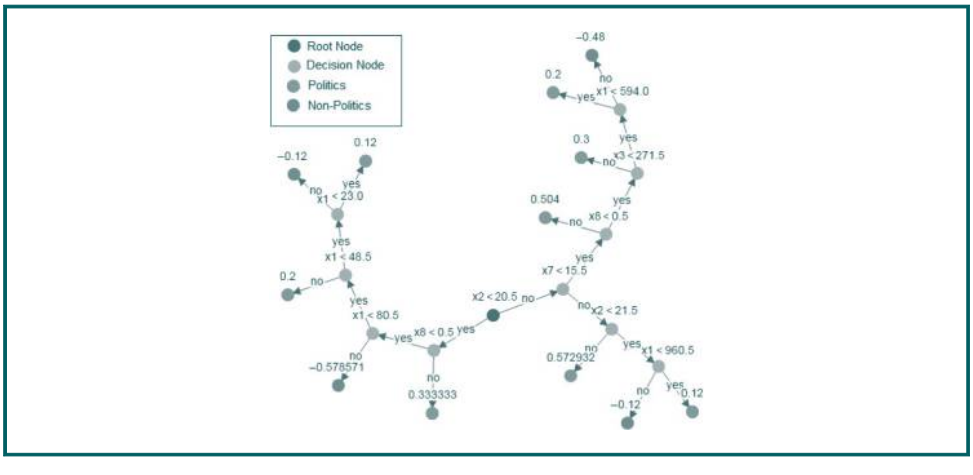
where μ_1 is the mean accuracy obtained by the LR and μ_2 is for the other methods, σ_1^2 is the accuracy variance obtained by the LR and σ_2^2 is for the other compared methods. N_1 is equal to 10 as this is a tenfold cross-validation. In general, the results indicate that there is no significant difference between LR and the other tested methods, although better accuracies can generally be obtained by the LR.

Therefore, the decision trees obtained by TD-DT can be interpreted and explained to executives of the user domain, as the accuracies obtained by TD-DT are similar to those obtained by LR. Figure 6 illustrates the resultant decision tree of the TD-DT classifier generated by Graphlab Create. It is evident that a feature is selected as a root node in TD-DT if this feature achieves the lowest classification error among the other features by

Table XI T-Values between LR and the other tested methods

	LR and SVM	*LR and TD-DR*	*LR and RF-DR*	*LR and GB-DR*
t-values	0.20842	1.1487	1.0703	0.99622

Figure 6 Decision tree created by TD-DT



applying the same data set. The values associated with each leaf node in Figure 6 represent the "margins" which are a form of prediction showing the distance of samples from the decision boundary. The greater the distance, the higher the confidence in the classifier's prediction that the user is interested in the politics domain. These margins can be converted to likelihood values (predictions) by applying the sigmoid function to the margins.

As depicted in Figure 6, Feature x_2 (fuchsia node) has been selected by the classifier as the root node at which to split the tree. To evaluate this tree, we start with the root node and follow the correct path through the decision nodes (green nodes) until we approach the leaf node (red/blue node), which indicates whether the user is interested in politics (red node) or not (blue node). For example, consider the two observations provided in Table XII; one indicates a user interested in politics (*@SenatorWacka*) and one does not show an interest in this domain (*@LabGallerie*). This Table shows the margins and the associated predictions for each sample. To apply the tree represented generated for *@SenatorWacka*, we start with the root node "$x_2 < 20.5$" is *no* because x_2SenatorWacka = 42, "$x_7 < 15.5$" is *no* because x_7SenatorWacka = 20, "$x_2 < 20.5$" is *no* because x_2SenatorWacka = 42. This leads to a red leaf with the value of "0.572932" which represents a user who is interested in the politics domain. The application of the same tree on *@LabGallerie* leads to a blue leaf with a value of −0.578571, which indicates a non-politics user. This is evident in both users, whose classification labels match with the resulting predictions.

5.2.1 A comparison with LDA and SLA. As discussed, LDA and SLA are statistically well-known models used for several topic modelling applications. In this section, we describe an experiment used to benchmark the applicability of our model at the user level against these two models, to identify a user's main topic of interest. Gensim's python implementation (Rehurek and Sojka, 2010) of LDA and SLA is used. The collected historical tweets of two politicians' accounts (i.e. @sarahinthesen8 and @stephenjonesALP) have been fed to the three models: LDA, SLA and our model incorporating a politics knowledge inference. The experimental settings for LDA and SLA are set to one topic modelling, and the extracted terms indicate the 25 most contributed terms to this topic. In our approach, we extract the top 25 frequently annotated entities from the users tweets. Tables XIII and XIV show the top 25 terms/entities extracted using the three approaches for @sarahinthesen8 and @stephenjonesALP, respectively.

The list of the top contributed terms identified using one-topic modelling for each user incorporating LDA and SLA illustrates the inadequacy of these approaches in identifying a high-level topic. On the other hand, with the top 25 entities annotated for both users using our approach, the high-level topic (i.e. politics) is highly noticeable. In our proposed system framework, each entity is linked with a specific class in the ontology. The knowledge obtained for each entity can be enriched to facilitate the overall semantic interlinking which leads to a better understanding of the domain of knowledge. Interlinking and enrichment are not applicable to LDA and SLA. Furthermore, all the top entities annotated using our proposed system framework indicate politics entities, although some of the most frequently occurring terms extracted using LDA and SLA are politics entities. In a nutshell, the

Table XII Margins and predictions of two samples

Twitter_ID	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*Label (1: politics, 0: non-politics)*	*Margins*	*Predictions*
@SenatorWacka	880	42	468	12	3	1	20	1	1	1	0.572932	0.639440
@LabGallerie	1498	4	19	1	2	7	3	0	0	0	−0.578571	0.359261

Table XIII Top entities/terms extracted using LDA, SLA and our approach for @sarahinthesen8

LDA	*LSA*	*Politics knowledge inference* *Entity*	*Sub-type*
refuge	refuge	Government of Australia	Organization
young	young	Australian Greens	Political Party
sarah	sarah	Member of parliament	Politician
hanson	hanson	Elections	Event
nauru	nauru	Australian Labor Party	Political Party
children	children	Parliament	Organization
detent	detent	Liberal Party of Australia	Political Party
govt	govt	Malcolm Turnbull	Politician
australia	australia	Peter Dutton	Politician
green	green	Tony Abbott	Politician
abbott	abbott	Politics	Ontology
today	today	Sarah Hanson-Young	Politician
asylum	asylum	Electorate	Voter
manu	manu	Council	Organization
aust	aust	Politician	Person
people	people	inequality	Political_Slogan
senate	senate	Coalition	Political_Slogan
seeker	seeker	Joe Hockey	Politician
abuse	abuse	George Brandis	Politician
news	news	Liberal National Party of Queensland	Political Party
minister	time	welfare	Political_Slogan
time	minister	Barnaby Joyce	Politician
dutton	dutton	Nick McKim	Politician
turnbull	turnbull	Kristina Keneally	Politician
australian	australian	Simon Birmingham	Politician

Table XIV Top entities/terms extracted using LDA, SLA and our approach for @stephenjonesALP

LDA	*LSA*	*Politics knowledge inference* *Entity*	*Sub-type*
illawarra	illawarra	Member of parliament	Politician
qt	qt	Elections	Event
today	today	Parliament	Organisation
great	great	Australian Labor Party	Political Party
mp	mp	Government of Australia	Organisation
stephen	stephen	Liberal Party of Australia	Political Party
good	good	Coalition	Political Slogan
post	post	Tony Abbott	Politician
school	school	Council	Organisation
abbott	abbott	Anthony Albanese	Politician
jone	jone	Politics	Ontology
photo	photo	Julia Gillard	Politician
auspol	auspol	Electorate	Voter
parliament	day	Greg Combet	Politician
day	jame	Sharon Bird	Politician
jame	parliament	Joe Hockey	Politician
big	big	Mark Butler	Politician
support	support	Malcolm Turnbull	Politician
nbn	nbn	Kate Ellis	Politician
house	house	Barack Obama	Politician
facebook	facebook	Joel Fitzgibbon	Politician
time	time	Jamie Briggs	Politician
fb	fb	Australian Greens	Political Party
australia	australia	Steven Ciobo	Politician
purser	purser	Greg Hunt	Politician

outcome of this experiment shows the applicability and effectiveness of our proposed framework.

5.3 Domain detection – tweet level

Table XV shows the statistics of the data set used for this experiment at the tweet level. The new tweets are collected from the list of users indicated in the previous section. These tweets represent the new tweets posted after Quarter 2, 2016. Hence, the tweet-level experiments are conducted on the set of tweets which have not been included in the users historical tweets as discussed in the previous section.

The features shown in Table II are formulated for each tweet. *On-topic samples* in Table I represent the list of tweets labelled as politics tweets. *Off-topic samples* show the list of tweets labelled as non-political tweets. The ARD value is calculated for each feature. Table XV shows the statistics calculated for the ground truth which is used to classify tweets according to a particular domain. It is evident that the calculated ARD for the two mean values of x_5 is the smallest value because of the noticeable convergence of x_5 in both categories. This is because a user who has been classified as belonging to the politics domain does not necessarily post all of his/her future tweets in this domain. Likewise, a user who has been classified as a non-politically interested user may show an interest in this domain in future tweets. Nevertheless, x_5 is most likely to distinguish the ambiguous political entities annotated from the textual content of a tweet, thereby helping to accurately ascertain the tweet's domain. This will be discussed further in this section.

Because of the ability of the LR to detect the domain of interest at the user level, LR is further used to classify tweets in this phase with the same set of parameters listed in Table VII. To validate the efficiency of the proposed approach, tenfold cross validation is performed where the 480 samples are randomly split into a training data set (80 per cent) and a validation data set (20 per cent). To further validate the effectiveness of the proposed approach, another experiment was conducted which excluded user_pol_likelihood, x_5 from the feature sets. This is to measure the significance of this feature to predict a tweets domain. Table XVI shows the confusion table used to quantify the performance of the LR classifier in each experiment, where Exp. 1 refers to the first experiment conducted incorporating all features listed in Table II. Exp. 2 refers to the second experiment conducted on the same data set excluding x_5.

Table XV Data set statistics – tweet level

	On-topic samples	*Off-topic samples*	*ARD*
Total #Tweets	255	225	
Total #political_entities (x_1)	880	71	85.068
Total #words_count (x_2)	3,762	2,391	22.282
Average political_perc, (x_3)	0.249	0.033	76.596
Total #pol_entities_recent_quarter (x_4)	65,049	37,248	27.177
Average user_pol_likelihood, x_5	*0.638*	*0.563*	*6.245*

Table XVI Confusion table – tweet level

Target label	*Predicted label*	*Exp. 1*	*Exp. 2*
On-topic	On-topic	*58*	*56*
Off-topic	Off-topic	*42*	*39*
On-topic	Off-topic	*0*	*3*
Off-topic	On-topic	*0*	*2*

As depicted in the confusion matrix in Table XVI, Exp. 1 achieved better results than Exp. 2; incorporating all features including the past user's political prediction (x_5) leads to zero incorrect classifications. However, eliminating x_5 from the list of features results in 5 out of 100 incorrect classifications. This is confirmed by the comparison of the performance results of the two experiments illustrated in Table XVII.

Despite the convergence in each metric listed in Table XVII, the predicted likelihoods of the validation data set incorporating all features closely match the assigned labels.

Table XVIII shows the highest estimated coefficient values calculated for each tweets feature in each of the conducted experiments. It is evident that political_perc (x_3) obtained the highest coefficient value in Exp. 1 and Exp. 2. This is because of the impact of the tweets political weight, indicating the tweets domain. This feature is supported by considering the number of political entities (x_1) and the total number of words in the tweet, x_2. User_pol_likelihood (x_5) obtained the second highest estimated coefficient after conducting Exp. 2. This is because of the significance of incorporating former knowledge about the user's political interest in the process of predicting the domain of their future tweets.

Table XIX elucidates further the significance of incorporating x_5. Table XIX shows two real tweets of the ground truth data set; one is labelled "politics" and the other is labelled "non-politics", posted by two Twitter users (i.e. @tamaleaver, non-politics user, and @peterjblack, politics user). The list of features included in Table II is calculated for each tweet. As depicted in Table XIX, Features x_1, x_2 and x_3 obtained the same values for each tweet. This exacerbates the process of obtaining the correct domain by considering only the number of political entities and counting the words in each tweet. It is evident that Features x_4 and x_5 are important for identifying the tweets domain because of their significance for the classification task.

Assumption. It is argued that the annotated political entity of a tweet posted by a user who has already been predicted to be interested in the politics domain, and who has included a relatively large number of political entities annotated in their tweets, is likely to indicate an

Table XVII Performance comparison of two experiments – tweet level

	Accuracy	*Log_loss*	*Precision*	*Recall*	*F1_score*
Exp. 1	1	0.01	1	1	1
Exp. 2	0.95	0.072	0.949	1	0.957

Table XVIII Highest positive coefficients – tweet level

Exp. 1		*Exp. 2*	
Feature	*Value*	*Feature*	*Value*
political_perc, x_3	24.86	political_perc, x_3	25.126
user_pol_likelihood, x_5	12.095	political_entities, x_1	1.823
political_entities, x_1	5.409	words_count, x_2	0.825
words_count, x_2	0.623	pol_entities_recent_quarter, x_4	0.009

Table XIX Features extracted for two tweets posted by two users (politics and non-politics)

Twitterer	*Tweet*	x_1	x_2	x_3	x_4	x_5	*Label*
@tamaleaver	"Researching microcelebrity: methods, access and *labour*, Jonathan Mavroudis"	1	7	1/7	4	*0.019*	*non-politics*
@peterjblack	"*Labor* could support 'self-executing' same-sex marriage plebiscite"	1	7	1/7	927	*0.98*	*politics*

actual political concept. Likewise, a user who has not shown an interest in politics in the past, is not likely to indicate politics-related content in future tweets. This helps to eliminate the ambiguity for those entities which might have dissimilar meanings in several contexts. Moreover, this is applicable to all domains of knowledge.

Therefore, despite obtaining one political entity (Labour/Labor) for each tweet in Table XIX, these tweets convey two different messages which are unrelated in terms of context.

6. Discussion

The rapid growth of enterprise needs in conjunction with an increase in the volume of modern data repositories, and the nature of the data that can be stored, has made traditional statistical methods insufficient to meet all data analysis requirements. This has necessitated the development of advanced data analytics to extract useful knowledge from such vast data volumes.

In light of the general perception of advanced data analytics, companies incorporate advanced social data analytics to build effective marketing strategies, leveraging the interactive quality of OSNs. Thus, to create the required interaction with their customers, companies use many modern forms of communication to attract customers and visitors to their online social platforms. Consequently, it is necessary for companies to analyse the customers' social content and classify the customers into appropriate categories, to deliver the right message to the right category. Segmentation is the first step towards effective marketing to classify customers according to their interests, needs, geographical locations, purchasing habits, life style, financial status and level of brand interaction. If companies succeed in building effective clusters of customers, and thus determine the basic criteria for each cluster in making their buying decisions, they will be able to take clear actions to implement them. For example, companies can identify the most optimal products/services captured for each segment of customers. This fine-grained analysis leads to maximisation of a customer's satisfaction with a company through designing and manufacturing several, segments-oriented products.

Unstructured data are produced exponentially. This necessitates further efforts to absorb such data sets to understand its context. Textual content (a.k.a natural language text) is considered the largest amongst all sources of information (Gupta and Lehal, 2009). The wealth of free-form textual social data has attracted the attention of researchers in an attempt to disclose hidden knowledge regarding textual content. This problem has been untangled through the emergence of text mining technology, an extension of data mining, which aims to detect rules, patterns and trends from textual data such as tweets, HTML webpages, instant messages and emails (Feldman and Sanger, 2007; He *et al.*, 2013). Natural language text is very ambiguous, and this is evident particularly when it comes to the continuous occurrences of the named entities. Hence, indicating and inferring key entities such as a person's name and profession, location of cities and countries, products, companies and specialised terms from the text can significantly enhance several business processes and techniques, such as knowledge base population, topics distillation, keyword searches and information integration (Shen *et al.*, 2015). Therefore, there is a need for an approach to derive knowledge from social big data. This approach enhances the overall comprehension of the processed textual data sets, and delivers knowledge in the form of unambiguous results through providing metadata which aids in accurately interpreting and understanding the related data.

Twitter is designed to track public figures and news, and provide a platform for users to follow their friends and associates. The "maximum 140 characters" quality has made Twitter particularly important and widespread; however, this feature constricts the size of published content for each user which is needed to conduct an adequate analysis. This paper presents an effective approach to address two main related problems:

1. the sporadic quality of tweets which entangles bag-of-words statistical techniques; and
2. the problematic nature of obtaining a factual understanding of the contextual meaning of users social content.

The most well-known approaches for inferring a user's topics of interest are the LDA-related techniques. Despite their popularity, they fail to address the following key issues:

- The number of topics to be discovered is set as a parameter in the experiment, thus it is hard to identify the optimal number which represents the adequate number of topics extracted from the document (Zhang *et al.*, 2017).
- The topics extracted by these models do not contemplate the temporal aspects. A document's corpus evolves through time and subsequently so does its themes (Alghamdi and Alfalqi, 2015).
- These models are considered as monolingual topic models, hence these do not differentiate idioms of the same language (Zoghbi *et al.*, 2016); and
- These models are unable to infer topics from short text, such as tweets (Li *et al.*, 2016).

Incorporating ontologies, semantic Web and linked data enriches textual data and the extraction of knowledge, thereby linking the textual data with a particular user domain. This approach is better able to address the temporal factor, and at harnessing advanced machine learning techniques to perform domain-based classification. For example, by recurrence to the benchmark comparison conducted in Section 5.2.1, if a user is interested in finding Twitter users who discuss "Australian political parties", through implementing an LDA technique, this user could find "Sarah Hanson-Young@sarahinthesen8" and "Stephen Jones@stephenjonesALP" amongst the search results. This is possible only if "Sarah" and "Stephen" indicated "Australian political parties" explicitly in their content alongside tweets pointing out their declared political party (i.e. "Australian Greens" and "Australian Labor Party", respectively).

Moreover, LDA retrieves search results that neglect the temporal dimension; users knowledge evolves over time and their interest might be diverted elsewhere depending on their experience, work, study or other factors. Leveraging domain ontology and semantic Web tools facilitates the building of conceptual hierarchies and the process of populating the domain ontology with instances extracted from user tweets. Therefore, "Australian Greens" and "Australian Labor Party" are annotated in the knowledge base as a subset of "Australian political parties". This hierarchy extends the knowledge obtained from social data by adding semantics to its textual content. Unlike LDA and other unsupervised statistical approaches, we incorporate supervised machine learning techniques to perform the classification task for the already semantically enriched temporally-segmented textual content. This, as indicated in the conducted experiments, validates the applicability of veritably classifying users based on their domains of interest which has an intrinsic impact on several applications. For example, adding a user-domain dimension when calculating trust in social media helps to provide a fine-grained trust analysis. In this context, the notion of domain-based trust for the data extracted from the unstructured content (such as social media data) is significant. This is achieved by calculating trustworthiness values which correspond to a particular user in a particular domain. This issue will be addressed in our future work as indicated in the following section.

7. Conclusion and future work

This paper presents the preliminary stages of a research project intended to provide a methodology for social business intelligence incorporating the notion of trust, semantic Web analysis and machine learning applications (Abu-Salih *et al.*, 2015). The importance of trust in the context of OSNs is indicated by the numerous resources available for market analysis,

listening to the VoC, and by sentiment analysis – all of which are major resources that feed business intelligence applications.

The semantic extraction of the textual content of OSNs represents a further step towards understanding the factual context of a user's content. One of the major challenges of OSN analysis is to better understand the domain of knowledge in which the user is interested. This problem is exacerbated by:

- inconsistent user behaviour (a user's interest can evolve and change over time) and
- the brevity and economy of tweets' content.

Therefore, this paper presents a consolidated approach to addressing this problem by means of semantic analysis and the application of machine learning.

The proposed framework comprises two analysis phases:

1. The time-aware semantic analysis of users historical content incorporating five well-known machine learning classifiers. This classifies users into two main categories: politics-interested and non-politics-interested.
2. The prediction likelihood values obtained in the first phase have been harnessed to predict the domain of the users future tweets.

The experiments conducted to evaluate this framework validate the applicability and effectiveness of better understanding the domain of Twitter content at the user and tweet levels. This is evident through the notable performance of the machine learning experiments conducted at both the user and tweet levels.

Through experiments conducted using the Twitter platform as one of the dominant OSNs, this work provides the essential groundwork for a better understanding of user interest in several domains of knowledge. This is achieved by incorporating domain-based ontologies and semantic Web analysis to gain a better familiarity with user interests. This facilitates the process of measuring user credibility in each domain of knowledge. The following are the possible enhancements and research directions to be addressed in our anticipated future work:

- Beside politics, a domain-based analysis of several domains of knowledge will be conducted to gain a more comprehensive insight into each domain. This is to facilitate the development of several domain-based ontologies leveraged by semantic Web technologies and linked open data.
- A domain-based trustworthiness approach will be developed based on the factual understanding of the users main interests.
- Machine learning will be harnessed to achieve the abovementioned research objectives through multi-classification applications, to predict the likelihood of user interest in several domains of knowledge.
- Semantic analysis and trust will be integrated for social business intelligence applications, which will enhance the quality and accuracy of data stored in data warehouses. This will dramatically affect the decision-making process as well as the quality of extracted reports.

Notes

1. AlchemyAPI is accuired by IBM's Watson since 2015.
2. http://www.gartner.com/document/3383817?ref=solrAll&refval=175496307&qid=34ddf525422cc71383ee22c858f2238a, Visited in 25/10/2016.
3. https://dev.twitter.com/rest/reference/get/statuses/user_timeline.

4. This threshold is set by Twitter™ as the maximum number of recent tweets the twitter API is allowed to retrieve.
5. https://dev.twitter.com/rest/reference/get/statuses/lookup.
6. https://wordnet.princeton.edu/
7. http://www.alchemyapi.com/products/alchemylanguage/linked-data.
8. https://blog.twitter.com/2016/announcing-an-application-process-for-verified-accounts-0.
9. http://www.aph.gov.au/.
10. https://twitter.com/latikambourke/lists/australian-journalists/subscribers, available at: https://twitter.com/lizziepops/lists/politics/members; https://twitter.com/smh/lists/federal-politicians
11. http://earleyedition.com/2009/04/22/australias-top-100-journalists-and-news-media-people-on-twitter; Wikipedia: Australian political journalists, available at:https://en.wikipedia.org/wiki/Category:Australian_political_journalists

References

Abu-Salih, B., Wongthongtham, P. and Zhu, D. (2015), "A preliminary approach to domain-based evaluation of users' trustworthiness in online social networks", *IEEE International Congress on Big Data*, pp. 460-466.

Abu-Salih, B., Wongthongtham, P., Beheshti, S. and Zajabbari, B. (2015), "Towards a methodology for social business intelligence in the era of big social data incorporating trust and semantic analysis," *Second International Conference on Advanced Data and Information Engineering (DaEng-2015), Springer, Bali*.

Alam, M.H., Ryu, W.J. and Lee, S. (2017), "Hashtag-based topic evolution in social media", World Wide Web, pp. 1-23.

Alghamdi, R. and Alfalqi, K. (2015), "A Survey of Topic Modeling in Text Mining".

Al-Tahrawi, M. (2015), "Arabic text categorization using logistic regression", International Journal of Intelligent Systems and Applications, Vol. 7, pp. 71-78.

Altınel, B., Can Ganiz, M. and Diri, B. (2015), "A corpus-based semantic kernel for text classification by using meaning values of terms", *Engineering Applications of Artificial Intelligence*, Vol. 43, pp. 54-66.

Anthes, G. (2010), "Topic models vs. unstructured data", *Communications of the ACM*, Vol. 53 No. 12, pp. 16-18.

Ardichvili, A., Maurer, M., Li, W., Wentling, T. and Stuedemann, R. (2006), "Cultural influences on knowledge sharing through online communities of practice", *Journal of knowledge management*, Vol. 10 No. 1, pp. 94-107.

Asharaf, S. and Alessandro, Z. (2015), "Generating and visualizing topic hierarchies from microblogs: an iterative latent dirichlet allocation approach", *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pp. 824-828.

BBC. (2014), "BBC Politics Ontology", 21 September 2016, available at: www.bbc.co.uk/ontologies/politics

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The semantic web", *Scientific american*, Vol. 284 No. 5, pp. 28-37.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of machine Learning Research*, Vol. 3, pp. 993-1022.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *Journal of machine Learning Research*, Vol. 3, pp. 993-1022.

Bolotaeva, V. and Cata, T. (2010), "Marketing opportunities with social networks", *Journal of Internet Social Networking and Virtual Communities*, Vol. 2010, pp. 1-8.

Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992), "A training algorithm for optimal margin classifiers", *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152.

Box, G.E., Hunter, J.S and Hunter, W.G. (2005), *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley-Interscience, New York, NY, Vol. 2.

Caruana, R. and Niculescu-Mizil, A. (2006), "An empirical comparison of supervised learning algorithms", *Proceedings of the 23rd International Conference on Machine Learning*, pp. 161-168.

Chen, M., Mao, S., Zhang, Y. and Leung, V.M. (2014), "Open issues and outlook," *Big Data*, Springer International Publishing, New York, NY, pp. 81-89.

Chen, Y., Yu, B., Zhang, X. and Yu, Y. (2016), "Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals", *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 1-5.

Chianese, A., Marulli, F. and Piccialli, F. (2016), "Cultural Heritage and Social Pulse: a Semantic Approach for CH Sensitivity Discovery in Social Media Data", *IEEE Tenth International Conference on Semantic Computing (ICSC)*.

De Maio, C., Fenza, G., Gallo, M., Loia, V. and Parente, M. (2017), "Time-aware adaptive tweets ranking through deep learning", Future Generation Computer Systems, 25 July.

De Maio, C., Fenza, G., Loia, V. and Orciuoli, F. (2017), "Unfolding social content evolution along time and semantics", *Future Generation Computer Systems*, Vol. 66, pp. 146-159.

De Nart, D., Degl'Innocenti, D., Basaldella, M., Agosti, M. and Tasso, C. (2015), "A Content-Based Approach to Social Network Analysis: a Case Study on Research Communities", Digital Libraries on the Move: *11th Italian Research Conference on Digital Libraries, IRCDL, Bolzano (29-30 January), Revised Selected Papers 2016, in Calvanese, D., De Nart, D. and Tasso, C. (Eds.), Springer International Publishing, Cham*, pp. 142-154.

Demchenko, Y., Grosso, P., De Laat, C., and., and Membrey, P. (2013), "Addressing big data issues in scientific data infrastructure", *Collaboration Technologies and Systems (CTS), International Conference on*, pp. 48-55.

Dong, L., Feng, N., Quan, P., Kong, G., Chen, X. and Zhang, Q. (2016), "Optimal kernel choice for domain adaption learning", *Engineering Applications of Artificial Intelligence*, Vol. 51, pp. 163-170.

Duggan, M. (2016), "The Political Environment on Social Media", 15 September 2017, available at: www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

Feldman, R. and Sanger, J. (2007), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge university press, Cambridge.

Friedman, J.H. (2001), "Greedy function approximation: a gradient boosting machine", *Annals of statistics*, Vol. 29 No. 5, pp. 1189-1232.

Ghahremanlou, L., Sherchan, W. and Thom, J.A. (2014), "Geotagging twitter messages in crisis management", *The Computer Journal*, Vol. 58, p. bxu034.

Gruber, T.R. (1995), "Toward principles for the design of ontologies used for knowledge sharing?", *International journal of human-computer studies*, Vol. 43 Nos 5/6, pp. 907-928.

Gupta, V. and Lehal, G.S. (2009), "A survey of text mining techniques and applications", *Journal of Emerging Technologies in Web Intelligence*, Vol. 1 No. 1, pp. 60-76.

Halberstam, Y. and Knight, B. (2016), "Homophily, group size, and the diffusion of political information in social networks: evidence from Twitter", *Journal of Public Economics*, Vol. 143, pp. 73-88.

Hassan, M.M., Karray, F., and., and Kamel, M.S. (2012), "Automatic Document Topic Identification using Wikipedia Hierarchical Ontology", *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pp. 237-242.

He, W., Zha, S. and Li, L. (2013), "Social media competitive analysis and text mining: a case study in the pizza industry", *International Journal of Information Management*, Vol. 33 No. 3, pp. 464-472.

Herzig, J., Mass, Y., and., and Roitman, H. (2014), "An author-reader influence model for detecting topic-based influencers in social media", *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pp. 46-55.

Ho, T.K. (1995), "Random decision forests", *Document Analysis and Recognition, Proceedings of the Third International Conference on*, pp. 278-282.

Hofmann, T. (1999), "Probabilistic latent semantic indexing", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50-57.

Hosmer, D.W., Jr, Lemeshow, S. and Sturdivant, R.X. (2013), *Applied Logistic Regression Vol. 398*, John Wiley & Sons, New York, NY.

Ito, J., Song, J., Toda, H., Koike, Y., and., and Oyama, S. (2015), "Assessment of tweet credibility with LDA features", *Proceedings of the 24th International Conference on World Wide Web*, pp. 953-958.

Iwanaga, I.S.M., The-Minh, N., Kawamura, T., Nakagawa, H., Tahara, Y. and Ohsuga, A. (2011), "Building an earthquake evacuation ontology from twitter", *IEEE International Conference on Granular Computing (GrC)*, pp. 306-311.

Kaplan, A.M. and Haenlein, M. (2010), "Users of the world, unite! the challenges and opportunities of social media", *Business Horizons*, Vol. 53 No. 1, pp. 59-68.

Karami, A., Gangopadhyay, A., Zhou, B., and., and Kharrazi, H. (2017), "Fuzzy Approach Topic Discovery in Health and Medical Corpora", International Journal of Fuzzy Systems, 1, pp. 1-12.

Kaushik, R., Apoorva Chandra, S., Mallya, D., Chaitanya, J.N.V.K. and Kamath, S.S. (2015), "Sociopedia: an Interactive System for Event Detection and Trend Analysis for Twitter Data", in Nagar, A., Mohapatra, D.P. and Chaki, N. (Eds), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics: ICACNI*, Volume 2, *Springer, New Delhi*, pp. 63-70.

Kumar, A. and Joshi, A. (2017), "Ontology Driven Sentiment Analysis on Social Web for Government Intelligence", Presented at the Proceedings of the Special Collection on eGovernment Innovations in India, New Delhi.

Lavbič, D., Žitnik, S., Šubelj, L., Kumer, A., Zrnec, A. and Bajec, M. (2012/2013), "Traversal and relations discovery among business entities and people using semantic web technologies and trust management", *Databases and Information Systems VII: Selected Papers from the Tenth International Baltic Conference, DB&IS*, p. 164.

Lee, K., Caverlee, J. and Webb, S. (2010), "Uncovering social spammers: social honeypots + machine learning", Presented at the Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva.

Li, C., Wang, H., Zhang, Z., Sun, A. and Ma, Z. (2016), "Topic Modeling for Short Texts with Auxiliary Word Embeddings", *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 165-174.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R. and Roxburgh, C. (2011), *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, London.

McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001), "Birds of a feather: homophily in social networks", *Annual Review of Sociology*, Vol. 27 No. 1, pp. 415-444.

Mehta, Y. and Buch, S. (2016), "Semantic proximity with linked open data: a concept for social media analytics", *International Conference on Computing, Communication and Automation (ICCCA)*,pp. 337-341.

Michelson, M. and Macskassy, S.A. (2010), "Discovering users' topics of interest on twitter: a first look", *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pp. 73-80.

Nichols, L.G. (2014), "A topic model approach to measuring interdisciplinarity at the national science foundation", *Scientometrics*, Vol. 100 No. 3, pp. 741-754.

Onan, A., Korukoglu, S. and Bulut, H. (2016), "LDA-based topic modelling in text sentiment classification: an empirical analysis", *International Journal of Computational Linguistics and Applications*, Vol. 7, pp. 101-119.

Quercia, D., Askham, H. and Crowcroft, J. (2012), "TweetLDA: supervised topic classification and link prediction in Twitter", presented at the the 4th Annual ACM Web Science Conference, Evanston, Illinois.

Quinlan, J.R. (1993), *C4. 5: Programming for Machine Learning*, Morgan Kauffmann, San Francisco, CA, p. 38.

Rainie, L. and Wellman, B. (2012), *Networked: The New Social Operating System*, Mit Press, Cambridge.

Rehurek, R. and Sojka, P. (2010), "Software framework for topic modelling with large corpora", *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Romero, S. and Becker, K. (2017), "Improving the classification of events in tweets using semantic enrichment", Presented at the Proceedings of the International Conference on Web Intelligence, Leipzig.

Russell, J.A. (2003), "Core affect and the psychological construction of emotion", *Psychological Review*, Vol. 110 No. 1, pp. 145-172.

Saif, H., He, Y., Fernandez, M. and Alani, H. (2016), "Contextual semantics for sentiment analysis of twitter", *Information Processing & Management*, Vol. 52 No. 1, pp. 5-19.

Saif, H., Dickinson, T., Kastler, L., Fernandez, M. and Alani, H. (2017), "A Semantic Graph-Based Approach for Radicalisation Detection on Social Media", The Semantic Web: 14th International Conference, ESWC, *Portorož, Slovenia*, (28 May– 1 June), Proceedings, Part I, Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P. and Hartig, O. (Eds), Springer International Publishing, Cham, pp. 571-587.

Salathé, M., Vu, D., Khandelwal, S. and Hunter, D. (2013), "The dynamics of health behavior sentiments on a large online social network", *EPJ Data Science*, Vol. 2 No. 1,

Sayce, D. (2016), "10 Billions Tweets… number of tweets per day", available at: www.dsayce.com/social-media/10-billions-tweets/

Schonhofen, P. (2006), "Identifying Document Topics Using the Wikipedia Category Network", presented at the Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

Scuotto, V., Del Giudice, M. and Carayannis, E.G. (2017), "The effect of social networking sites and absorptive capacity on SMES'innovation performance", *The Journal of Technology Transfer*, Vol. 42 No. 2, pp. 409-424.

Scuotto, V., Del Giudice, M. and Obi Omeihe, K. (2017), "SMEs and mass collaborative knowledge management: toward understanding the role of social media networks", *Information Systems Management*, Vol. 34 No. 3, pp. 280-290.

Scuotto, V., Del Giudice, M., Peruta, M.R.D. and Tarba, S. (2017), "The performance implications of leveraging internal innovation through social media networks: an empirical verification of the smart fashion industry", *Technological Forecasting and Social Change*, Vol. 120, pp. 184-194.

Sendi, M., Omri, M.N. and Abed, M. (2017), "Possibilistic interest discovery from uncertain information in social networks", *Intelligent Data Analysis*, Vol. 21 No. 6.

Shapiro, M.A. and Hemphill, L. (2017), "Politicians and the policy agenda: does use of twitter by the US congress direct New York times conten?", *Policy & Internet*, Vol. 9 No. 1, pp. 109-132.

Sharef, N.M., Martin, T., Kasmiran, K.A., Mustapha, A., Sulaiman, M.N. and Azmi-Murad, M.A. (2015), "A comparative study of evolving fuzzy grammar and machine learning techniques for text categorization", *Soft Computing*, Vol. 19 No. 6, pp. 1701-1714.

Shen, W., Wang, J. and Han, J. (2015), "Entity linking with a knowledge base: issues, techniques, and solutions", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27 No. 2, pp. 443-460.

Sherchan, W., Nepal, S. and Paris, C. (2013), "A survey of trust in social networks", *ACM Computing Surveys*, Vol. 45 No. 4.

Smith, M., Szongott, C., Henne, B. and Voigt, G.V. (2012), "Big data privacy issues in public social media", *6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, pp. 1-6.

Tess, P.A. (2013), "The role of social media in higher education classes (real and virtual) – a literature review", *Computers in Human Behavior*, Vol. 29 No No. 5, pp. A60-A68.

Van Kessel, S. and Castelein, R. (2016), "Shifting the blame. Populist politicians' use of Twitter as a tool of opposition".

Vicient, C. and Moreno, A. (2015), "Unsupervised topic discovery in micro-blogging networks", *Expert Systems with Applications*, Vol. 42 Nos 17/18, pp. 6472-6485.

Wang, C., Thiesson, B., Meek, C. and Blei, D. (2009), "Markov topic models", *Artificial Intelligence and Statistics*, *Columbia*, pp. 583-590.

Weng, J., Lim, E.-P., Jiang, J. and He, Q. (2010), "Twitterrank: finding topic-sensitive influential twitterers", *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 261-270.

Xiao, C., Zhang, Y., Zeng, X. and Wu, Y. (2013), "Predicting user influence in social media", *JNW*, Vol. 8 No. 11, pp. 2649-2655.

Yen, S.J., Lee, Y.S., Ying, J.C. and Wu, Y.C. (2011), "A logistic regression-based smoothing method for Chinese text categorization", *Expert Systems with Applications*, Vol. 38, pp. 11581-11590.

Zhang, W., Cui, Y. and Yoshida, T. (2017), "En-LDA: an novel approach to automatic bug report assignment with entropy optimized latent dirichlet allocation", *Entropy*, Vol. 19 No. 5, p. 173.

Zoghbi, S., Vulić, I. and Moens, M.-F. (2016), "Latent dirichlet allocation for linking user-generated content and e-commerce data", *Information Sciences*, Vol. 367-368, pp. 573-599.

Corresponding author

Pornpit Wongthongtham can be contacted at: ponnie.clark@curtin.edu.au