



Relational Learning Analysis of Social Politics using Knowledge Graph Embedding

Bilal Abu-Salih, et al. *[full author details at the end of the article]*

Received: 12 November 2020 / Accepted: 27 April 2021 / Published online: 12 May 2021

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2021

Abstract

Knowledge Graphs (KGs) have gained considerable attention recently from both academia and industry. In fact, incorporating graph technology and the copious of various graph datasets have led the research community to build sophisticated graph analytics tools, which has extended the application of KGs to tackle a plethora of real-life problems in dissimilar domains. Despite the abundance of the currently proliferated generic KGs, there is a vital need to construct domain-specific KGs. Further, quality and credibility should be assimilated in the process of constructing and augmenting KGs, particularly those propagated from mixed-quality resources such as social media data. For example, the amount of the political discourses in social media is overwhelming yet can be hijacked and misused by spammers to spread misinformation and false news. This paper presents a novel credibility domain-based KG Embedding framework. This framework involves capturing a fusion of data related to politics domain and obtained from heterogeneous resources into a formal KG representation depicted by a politics domain ontology. The proposed approach makes use of various knowledge-based repositories to enrich the semantics of the textual contents, thereby facilitating the interoperability of information. The proposed framework also embodies a domain-based social credibility module to ensure data quality and trustworthiness. The utility of the proposed framework is verified by means of experiments conducted on two constructed KGs. The KGs are then embedded in low-dimensional semantically-continuous space using several embedding techniques. The effectiveness of embedding techniques and social credibility module is further demonstrated and substantiated on link prediction, clustering, and visualisation tasks.

Keywords Domain-specific Knowledge Graphs · Social Politics · Knowledge Graph Embedding · Knowledge Graph Completion · Knowledge Graph Construction · Semantic Analytics · Trustworthy Knowledge Graphs

Communicated by Tim Weninger.

1 Introduction

Knowledge Graphs (KGs) have now been widely incorporated in industry and academia as being a factual reflection of the human knowledge to solve several domain-dependent real-life problems (Ji et al. 2020). In fact, the proliferation of Social Big Data has prompted the necessity for sophisticated approaches to assist machines to better understand the context of the multimodal contents. In particular, the heterogeneity in data sources and formats, the discrepancy in vocabulary, and the lack of comprehensive and integrated knowledge repositories are the key challenges for analysts. Yet, by presenting the domain knowledge as a set of entities and relations, KGs facilitate constructing a unified standard representation for the fusion of data. This thereby has led to knowledge propagation embodying graph datasets of divergent and interrelated domains (Wang et al. 2018a, b) and has extended to benefit large scale applications such as question answering (Zhang et al. 2018), recommender systems (Palumbo et al. 2017), KG completion (Lin et al. 2015), entity disambiguation (Huang et al. 2015a, b), and text classification (Marin et al. 2014). As a result, analysts today are able to conduct an in-depth analysis of external business data such as customer blog postings (Gruhl et al. 2004), Internet chain-letter data (Liben-Nowell and Kleinberg 2008), social tagging (Anagnostopoulos et al. 2008), Facebook news feed (Sun et al. 2009), and many other semantic Artificial Intelligence applications.

Despite the widespread usage of domain-independent (open-world) KGs (e.g. Google KG), domain-dependent KGs provide an overabundance of benefits to tackle domain-specific problems as well as to gain the hoped-for added value from domain corpora (Kejriwal et al. 2019). Domain knowledge is commonly captured in a KG, which is then used to enrich the semantics of data with a specific conceptual representation of entities. The reuse of domain ontology and interlinking process of embodied classes, entities, and concepts with other relevant entities from other KG repositories, facilitates the interoperability of information. Hence, KGs are used as backbones to support intelligent systems by extracting the semantics of textual data which are collected from different vocabularies and semantic repositories to enrich the semantic description of resources using an annotation component.

Another important consideration is the factuality and credibility of the embodied knowledge in a KG. The rapid growth in KGs sizes has risen a question on the quality of the embodied knowledge (i.e. entities and relations), and whether these facts do factually represent the intended real-world entities interlinked via their relationships. This has posed several research challenges in this field. For the purpose of proof of concept, this study targets the social political domain due to the fact that social media has been broadly used as an important arena by politicians to promote campaigns and to express and defend their views, and to open direct dialogues with their supporters (Shapiro and Hemphill 2017). Further, the amount of political discourse in social content is increasing; over 55% of OSNs' users believe that they are worn-out by political posts and discussions (Anderson and Quinn 2020). This propagation of political social contents yet can be hijacked

and misused by spammers to spread misinformation and false news. Hence, the data collected from OSNs should be scrutinised to augment KGs with trustworthy facts to benefit real-life applications. Despite the significant efforts attempted to address quality, endeavours in this direction are inadequate, and several measures should be proposed and taken to maintain the quality of KGs.

This paper presents a novel credibility-based domain-specific KG Embedding (KGE) framework. This framework comprises the following key modules: (1) Domain knowledge inference: involves capturing real-life entities obtained from social data into a formal and integrated representation depicted by a domain ontology (Politics). The proposed module makes use of various cross-domain knowledge-based repositories including Google KGTM, IBM Watson NLUTM, and WordnetTM to enrich the semantics of the textual contents, thereby facilitating the interoperability of information. (2) Social credibility: to measure the credibility of collected users from social media and their content, thereby eliminating spam and low trustworthy content from further analysis. This module incorporates several fine-grained key attributes to establish feature-based ranking model and reflect this model by means of their domain-based credibility values. (3) KG construction and embedding: the aim of this module is twofold: (a) to construct the KG based on the underlying abstract structure of Politics ontology and leveraging various mapping, annotation, enrichment and interlinking methods; and (b) to embed the constructed KG in low dimensional vector space using several embedding techniques.

The resultant KG embeddings of two separate KGs (original and curated) are used to conduct several tasks including link prediction, clustering, and visualisation. Evaluation protocol and metrics are used to compute the performance of the incorporated embedding models and to prove the effectiveness of the framework and the embodied modules.

In this paper, we have made the following key contributions:

- A domain knowledge graph is constructed based on an extended politics domain ontology using dissimilar light-weight ontologies and semantic repositories.
- An embedded social credibility module is incorporated and customised to enhance the quality of the collected datasets.
- Various state-of-the-art embedding models are implemented and their performance is evaluated using key evaluation metrics.
- The utility of the constructed KG Embeddings is demonstrated and substantiated on link prediction, clustering, and visualisation tasks.

This paper is organised as follows: Sect. 2 provides background on works related to the context of this paper. Section 3 discusses the overall methodology of the proposed framework of this paper and the included modules. The experiments carried out in this study are explained in Sect. 4 along with the evaluation mechanism and the implemented tasks. Finally, the conclusions and some possible research directions are reported in Sect. 5.

2 Background and related works

Domain-specific KGs Domain-specific/dependent KGs are constructed from domain corpora to establish relevant and semantically interrelated ground to tackle a specific domain problem. Therefore, domain-specific KGs can be defined as the process of enriching an underlying domain ontology (Kejriwal 2019). Also, it can be more comprehensively defined as an “*explicit conceptualisation to a specific subject-matter domain represented in terms of semantically interrelated entities and relations*” (Abu-Salih 2021). There have been continuous attempts to construct KGs to capture several domains of knowledge. For example, in the politics domain, (Nguyen and Jung 2019) created a KG that captured and cluster social events decomposed from social media using Independent Component Analysis (ICA). This was followed by using SocioScope Knowledge Graph (SKG) model to automatically construct event-driven KGs from Twitter data (Laufer and Schwabe 2017) presented POLARE, an ontology for political system conceptualisation. This ontology is then used to build a KG so as to be used for a better understanding of the existing relations between agents in the political system in Brazil. Capturing politics domain has been also addressed in Chen et al. (2017) and Huang et al. (2017). Constructing domain-specific KGs have been also extended to different domains, such as Healthcare (Cui et al. 2020; Sheng et al. 2020), Education (Chen et al. 2018; Zheng et al. 2017), ICT (Kiesling et al. 2019) (Deng et al. 2019), Sciences and Engineering (Gong et al. 2021; Liu et al. 2021), Finance (Tong et al. 2016) (Liu et al. 2019), and Travel and Tourism (Feng 2020; Liang et al. 2020; Wu et al. 2020).

Knowledge acquisition (completion, entity and relation extraction) The ongoing efforts to construct large-scale KGs have been notably increasing. This has led to producing massive KGs embodying billions of facts that describe different contexts (Rossi et al. 2020). These KGs, however, suffer from incompleteness, which negatively affects the utility of such graphs to be leveraged in real-life applications (Akrami et al. 2018). For example, Freebase, a large-scale KG and commonly used knowledge-base in research communities, is far from completeness; it has been indicated that the “place of birth” for above 70% of “Person” entities are missing, and more than 90% of the person entities have no embodied ethnicity (West et al. 2014; Dong et al. 2014). This also applies to Wikipedia and many other knowledge bases. This has led the research community to confront this issue by providing technical solutions to tackle it, commonly known as KG Augmentation/Completion approaches. KG Completion (a.k.a Link Prediction) aims to enrich the KG with new facts that are depicted by new likelihood entities and/or new relations. Link prediction has many applications, such as predicting new friendships in social networks and recommender systems to various other use cases. In this context, a new cohort of models has recently gained considerable attention. These models are designed to embed the constituents of a KG (entities and relationships) into a low-dimension semantically-continuous space (Wang et al. 2017a, b). The generated embeddings can be then leveraged to generate a set of candidate facts to fulfil a completion task (Meilicke et al. 2018).

KGs are commonly constructed from (semi-)structured (e.g., Wikipedia) or unstructured (e.g., web data) datasets. However, harvesting meaningful information from heterogeneous data sources is not a trivial task. It encompasses extracting facts (entities linked via relationships) that require a correlated array of various Information Extraction (IE) techniques, Natural Language Processing (NLP), and other statistical approaches (Paulheim 2017). Examples of techniques used for entities recognition and relations extraction are; Conditional Random Field (CRF) (Lin and Wu 2009), Machine Learning models (e.g. SVM), Neural Networks models, such as Bidirectional Long Short-Term Memory (BiLSTM) (Huang et al. 2015a, b), Hidden Markov Models (HMM) (Morwal et al. 2012), and off-the-shelf NLP tools (e.g. spaCy,¹ Stanford CoreNLP,² AllenNLP,³ IBM Watson NLU,⁴ etc.).

KG can be further extended and thereby its embedded knowledge can be augmented to include missing facts of the real world leveraging contextualised knowledge repositories (Beheshti et al. 2020, 2018). The process of knowledge acquisition can be categorised into two key dimensions, namely KG completion, and entity and relation extraction. KG completion aims to expand the current knowledge by accumulating more facts to the current state of the KG, while the latter dimension aims to infer new knowledge by predicting new relations and entities (Ji et al. 2020). Link and entity inference in the context of KGs is the process of amplifying the KG with new facts depicted by new entities and/or new relations.

Several approaches have been introduced to tackle this issue (Han et al. 2018; Purohit et al. 2019; Lin et al. 2015; Balažević et al. 2019; Balazevic et al. 2019) (Kazemi and Poole 2018). These attempts have also extended to address interrelated domains (Qiuyu and Fuhua 2020). (Han et al. 2018) proposed a joint representation learning framework to solve the complexity of the structure of semantic information by presenting a mutual attention mechanism, which can be used to highlight the important features by conjoining the textual content and the KG models. Augmenting knowledge in disaster situations has been also addressed in the literature. For example, (Purohit et al. 2019) proposed DisasterKG; a disaster KG that offers a platform that provides resources to answer critical inquiries. The authors made their point on how interoperability of information from dissimilar data resources can efficiently improve decision making in such cases. Completion of KG by using web pages was attempted by (Kruit et al. 2019). (Kruit et al. 2019) suggested a new approach for HTML table interpretations, where the row and column indicate an entity and an attribute respectively. By using the Probabilistic Graphical Model (PGM), the authors were able to infer new facts for KGs with dissimilar topologies. (Shi and Weninger 2018) proposed ConMask; an Open-World Knowledge Graph Completion system. This system is designed incorporating fully convolutional neural networks, and semantic averaging to be able to tackle the incompleteness of

¹ <https://spacy.io/>.

² <http://corenlp.run/>.

³ <https://allennlp.org/>.

⁴ <https://www.ibm.com/cloud/watson-natural-language-understanding>.

KG. The proposed system has proven ability to forecast relations including unseen entities.

NLP applications using KGE Incorporating graph technology together with the abundance of dissimilar graph datasets has assisted in building quite sophisticated graph analytics tools. Despite the effectiveness of the conventional graph analysis approaches, such as Graphx (Gonzalez et al. 2014), Gephi (Bastian et al. 2009), GraphLab (Low et al. 2012) to name a few, graph embedding has notably improved the efficiency of conducting graph analytics by converting the graph to a low semantic dimensional space, thus information can be represented as vectors leading to computational efficiency. Several efforts have been conducted to incorporate KG Embeddings to address numerous NLP challenges. For example, Yao et al. (2017) proposed a topic distillation approach embodying Latent Dirichlet Allocation (LDA) to improve document presentation in the semantic space. Authors of (Li et al. 2018) benefited from the architecture of a neural network and a constructed knowledge base to build Text Concept Vector (TCV) that can be used to infer high-level presentation of concepts from the textual content. KGs are also utilised in conjunction with deep learning models to distil knowledge for several applications, such as sentiment tasks (Song 2019), bilingual dictionary induction (Nakashole and Flauger 2017), fake news detection (Pan et al. 2018), recommender systems (Wang et al. 2019), and other miscellaneous applications (Long et al. 2020; Yang et al. 2016).

Classification and clustering using KGE Classification in the context of KGs is the task of determining, whether the entities/nodes, relations/edges, or the whole triple contained within the testing dataset are correct. This task can be perceived as a binary classification task involving class labelling to each entity, relation or triple. Underneath this broad classification umbrella, quite a few conducted literature reported attempts to efficient and reliable applications incorporating graph embedding (Kipf and Welling 2016; Wang et al. 2017a, b) and also using dissimilar embedding techniques such as TransR (Lin et al. 2015), HolE (Nickel et al. 2016) and ANALOGY (Liu et al. 2017). On the other hand, clustering is an unsupervised learning approach that aims to assemble similar entities in groups. Clustering can also be used to examine the efficiency of the approach used for KG embedding. Incorporating KGE boosts the traditional clustering algorithms by transforming the embedded components of the graph into vectors (Cai et al. 2018). Other unconventional approaches have been also presented in the literature. For example, (Tian et al. 2014) showed how utilising deep neural networks can improve KG clustering through mapping the similarity matrix of the input graph to the output graph embedding using the layer-wise pre-training scheme.

3 Methodology

3.1 Overall framework architecture

Figure 1 shows the proposed KGE framework. As depicted in the figure, the system comprises five core components, namely: Domain Knowledge Acquisition & Pre-processing; Domain Knowledge Inference; Knowledge Credibility Module;

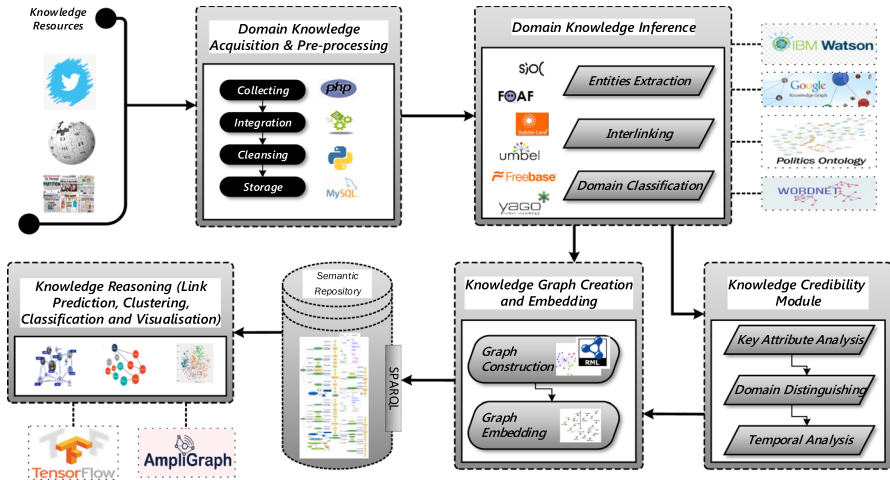


Fig. 1 System architecture

Knowledge Graph Creation and Embedding; and Knowledge Reasoning. The system collects its datasets from three main knowledge resources, namely: Twitter, Wikipedia, and miscellaneous news articles.

The collected datasets are pre-processed in order to ensure data cleansing and integration, then domain knowledge being captured in domain ontologies is identified and used to semantically enrich the textual contents. This process is attained through the domain knowledge inference module—semantic kitchen. This module incorporates several knowledge-based repositories including Google KG™, IBM Watson NLU™, Politics domain Ontology, and Wordnet™. The next phase in this framework is to ensure the credibility of the incorporated knowledge. The credibility of knowledge is commonly neglected in the construction of KGs especially when the knowledge is attained from social media where spammers and other low trustworthy users find a fertile medium to publish and spread their content taking advantage of the open environment and fewer restrictions of these platforms. The following module constructs the domain KG and conducts the KG embedding. This facilitates the knowledge reasoning which is carried out in the last module and represented by incorporated neural machine learning models for relational learning. Details on the system framework and the embodied modules are discussed in the next sections.

3.2 Domain knowledge acquisition and pre-processing

Since the emergence of the OSNs, the propagation of social data has revolutionised the research avenues to develop state-of-the-art techniques for social data analytics. OSNs are a fertile medium for researchers in diverse disciplines, leveraging the vast volume of content. For the purpose of proof of concept, this study focuses on analysing the political content that can be collected and distilled from the Twitter platform. The politics domain is selected amongst other domains due to the following reasons: (1) Twitter has been intensively incorporated as an important venue by

politicians to express and defend their policies, to practice electoral propaganda, and to communicate with their supporters (Shapiro and Hemphill 2017). (2) Twitter has raised a lot of controversy about its usage as a platform to attack political opponents (Van Kessel and Castelein 2016). (3) Twitter is characterized by its growing social base to include broad political social groups leveraged by ease of use, free to access, and less governmental control (Halberstam and Knight 2016). (4) In fact, the amount of political discourses amongst the overall social content is overwhelming; over 55% of OSNs' users believe that they are worn-out by political posts and discussions (Anderson and Quinn 2020). The social dataset used for this study has been collected using the twitter's "User_timeline"⁵ API method. This mechanism allows access to and retrieval of the public users' content and metadata.

3.2.1 Dataset acquisition

This study aims to augment the constructed domain-based knowledge graph with facts attained from heterogeneous data sources. These facts will not be only obtained from politics-related sources, but also be gathered from users who do not explicitly indicate an interest in this designated domain. Further, users who will be potentially detected as spammers will be also included to prove the applicability of our approach to filter out those users, thereby enhancing the quality of imported facts as will be discussed later.

Users who explicitly indicate an interest in politics domain are collected from various resources as follows: (1) we gathered all information provided for members listed in the Parliament of Australia official website.⁶ Those include Senators and members of the Australian House of Representatives. (2) A selected set of users is assembled from three distinguished Australian Twitter lists that are relevant to the political domain.⁷ (3) Mixt sources⁸: users whose political interest is not explicitly identified were tentatively selected various Australian Twitter's lists established to discuss sports, Information Technology, and other non-politics domains. (4) Finally, we included a subset of users indicated in the Twitter graph dataset collected by (Akcora et al. 2014). This graph was used in experiments carried out by Akcora et al. to discover spammers and other illegitimate accounts. One of the contributions of this paper is to provide a platform where trustworthy social content can be imported to augment the domain KG, and by this means eliminate untrustworthy content. Hence, the reason for selecting the graph of (Akcora et al. 2014) is twofold: (1) to proof the efficiency and applicability of the proposed approach which can be used to eliminate spammers and their content and entrench the domain KG with

⁵ https://dev.twitter.com/rest/reference/get/statuses/user_timeline.

⁶ <http://www.aph.gov.au/>.

⁷ <https://twitter.com/latikambourke/lists/australian-journalists/subscribers>; <https://twitter.com/lizziepops/lists/politics/members>; <https://twitter.com/smh/lists/federal-politicians>.

⁸ <http://earleyedition.com/2009/04/22/australias-top-100-journalists-and-news-media-people-on-twitter>; Wikipedia: Australian political journalists: https://en.wikipedia.org/wiki/Category:Australian_political_journalists.

trustworthy facts; (2) to embed also the content of domain influencers from a dataset of users whose domains of knowledge are not explicitly known.

3.2.2 Dataset pre-processing

One of the significant features of properly addressing and curating Big data is to ensure its veracity. The veracity of data refers to the certainty, faultlessness, and trustworthiness of data (Demchenko et al. 2013). Although reliability, availability, and security of data's source is significant (Demchenko et al. 2013), these factors do not guarantee data correctness and consistency especially in the context of social media where data can be infected with spam and other junk contents. Hence, appropriate data cleansing, integration, and credibility techniques should be incorporated to ensure the certainty and veracity of data. The collected users and their contents are cleansed and integrated to enhance quality as follows:

Datasets cleansing cleansing data is a crucial step to improve the quality of data that will be used in further analysis. Hence, detecting and removing errors and corrupted data, meaningless data, redundant data, and irrelevant data are key techniques in data cleansing which are carefully carried out in this experiment to guarantee that only curated data are passed for the next phase.

Data quality enhancement the list of Twitter handles (a.k.a. screen name such as @username), which are indicated in the user's metadata, is collected and replaced with the actual user's corresponding name. To achieve this task, Twitter provides a RESTful API service called "lookup"⁹ that is used to reveal the twitterer with a certain handle by receiving full hydrated information about the user. Twitter handles are commonly neglected in Twitter mining applications. However, handles are used to mention for example twitterers of important entities that are related to a certain domain. For example, a user demonstrates an interest in the political domain if the user is commonly posting politics-related content as well as mentioning twitterers related to politics domain such as politicians or political parties. Hence, it is essential to identify and determine the actual user information of those handles. This assists in the process of domain modelling and inference.

3.3 Domain knowledge modelling and inference

Domain knowledge modelling inference is the key phase in the proposed framework. Knowledge modelling presents the core activity in knowledge graph creation. It involves capturing the real-life entities obtained from the social data into a formal representation depicted by the domain ontology. Tom Gruber generated expansive interest across the computer science community by defining ontology as "an explicit specification of a conceptualisation" (Gruber 1993). While conceptualisation aims to formulate the knowledge about real-world entities, the specification attempts to represent those captured entities in a concrete form (Stevens 2001). Therefore, ontology

⁹ <https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-users-lookup>.

Fig. 2 BBC Politics ontology

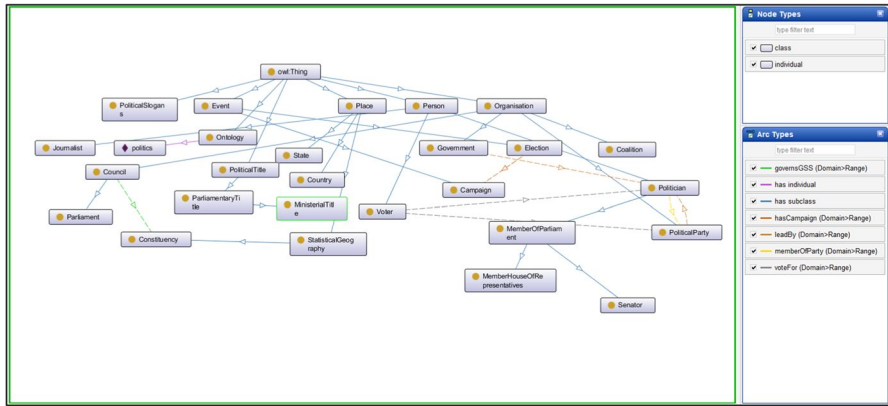
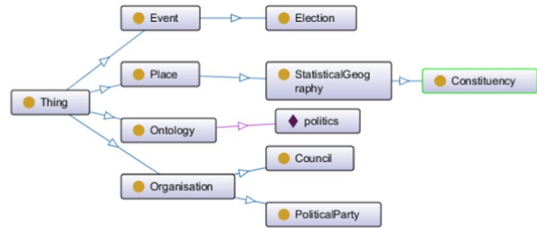


Fig. 3 BBC politics ontology extension

captures the domain knowledge through the defined concrete concepts (representing a set of entities), constraints, and the relationship between concepts, thereby providing a common understeering of the domain as well as giving a formal representation in machine-understandable semantics. The purpose of an ontology is to represent, share, and reuse existing domain knowledge. This module aims to detect and infer the user’s domain of knowledge from pre-processed datasets. For a proof of concept purpose, we experiment within the Politics domain. We use Politics ontology, WordNet, and ontology interoperability and integration to infer political knowledge.

Politics ontology The BBC offers an array of domain ontologies which are designed to conceptualise a predefined set of domains such as, sports, music, education, to cite a few (‘BBC Ontologies’ 2015). These domain ontologies are designed to consolidate the established BBC Linked Open Data platform. The politics domain is amongst the ontologies constructed by BBC and is described as the conceptual knowledge captured in politics ontology along with its embodying 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). Figure 2 displays the BBC Politics ontology. This ontology is initially designed to capture politics in the context of UK government elections.

However, the concepts and relationships embedded in the designated ontology are inadequate to properly model this nominated domain, particularly that this study addresses the domain of the politics in the Australian context. Hence, we extend the BBC politics ontology to provide a better depiction of the political domain. In this

study, the extension of the political ontology is conducted manually by one of the authors who is an expert in Australian politics domain. In particular, BBC Politics ontology is scrutinised and extended with further concepts/classes and relations to provide a better comprehension of this domain. In future, we aim to explore new venues for ontology augmentation using (semi-)automatic techniques. Figure 3 shows the extended version of the BBC Politics ontology which is used in this research.

Designing high-quality ontology is important as a corner store to provide a meaningful, contextualised, valid, error-free knowledge base. Also, the developed ontology for any domain should be appropriate to answer queries over its semantic concepts, relationships and instances. The new extended version of Politics ontology is therefore verified to ensure logical consistency. This has been carried out using the reasoning process. In particular, the extended ontology is reasoned incorporating various well-known reasoners such as TrOWL, RacerPro, Pellet, Pellet (Incremental), Hermit, FaCT+++. Besides the standard inference services provided by ontology reasoners such as classification and realisation, reasoners are generally used to scrutinise all concepts, properties, instances, and embedded hierarchies. Also, they check if concepts are satisfactory and their descriptions are free of contradiction. The new extended Politics Ontology is reasoned and verified, and no contradictory facts are indicated.

WordNet database WordNet¹⁰ is a vocabulary lexicon that includes a collection of terms/words (synsets /synonyms) that are interrelated and have similar semantic meanings. WordNet is commonly used to augment the term with further other semantic-related concepts that can enrich its meaning. For example, various synsets of the same contextual meaning can be extracted for the term ‘Pol, such as “*politician, politico, and political leader*”’. WordNet is used in this study to expand the knowledge base with synonyms to concepts inserted in the extended Politics Ontology.

Ontology interoperability Ontology interoperability aims to align and consolidate the developed ontology with relevant entities captured from other predefined domain and generic ontologies. Ontology interoperability is attained in this study by apprehending equivalent links (URIs) that indicate the same entity/resource. This linkage is depicted by using, for example, owl#sameAs relation for the resources in the Linked Data. This entails that URIs of both subject and object indicate the same resource. For the interlinking process, we incorporate Google KG, a knowledge base that is mainly developed to enhance Google’s search engines by providing relevant, semantically-enhanced, and context-specific results. The Google KG Search API¹¹ is used to infer entities and categorised classes/types. In particular, Google KG Search API provides a platform to collect entities that belong to a wide variety of independent domains. However, our approach incorporates a domain-driven approach that uses a domain ontology to find relevant entities captured from the textual content of users’ tweets. Therefore, amongst all entities that are obtained from analysing tweets

¹⁰ <https://wordnet.princeton.edu/>.

¹¹ <https://developers.google.com/knowledge-graph>.

using Google KG Search API, we incorporate those that are mapped with the concepts of our politics ontology. Hence, irrelevant entities which are not related to our designated domain are neglected.

We also utilised the Natural Language Understanding service of IBM Watson™ as a one-stop-shop, leveraging access to a wide variety of linked data resources through providing easy access APIs. These resources include but are not limited to: different vocabularies such as Upper Mapping and Binding Exchange Layer (UMBEL), Freebase which are community-curated databases for well-known people, places, and things, YAGO high-quality knowledge base, etc.

IBM Watson is also used for domain-based classification. In particular, IBM Watson analyses the given text or URL and categorises the content of the text or webpage according to a set of categories (taxonomies) with the corresponding scores. Scores are calculated using IBM Watson, range from “0” to “1”, and convey the precise degree of an assigned Category/Taxonomy/Domain to the processed text or webpage. IBM Watson presents an inclusive list of categories divided into certain predefined hierarchies where the high-level category indicates the high-level category and the deeper-level category provides a fine-grain category analysis. For instance, “law, govt and politics” is considered a high-level category in which “presidential elections” is one of its deep-level categories. IBM Watson is used further to identify the overall positive or negative sentiment of the provided content. The taxonomy inference module is used in this research in the domain discovery process, while sentiment analysis is used to discover the sentiments of tweets’ replies. The purpose of domain classification and sentiment analysis is discussed in the following section.

3.4 Social credibility module

As mentioned previously, this study aims to make use of domain-specific politics ontology and available KGs to analyse the social contents of users in OSNs, thereby augmenting the domain KG with facts inferred from users with legitimate and credible interest in politics domain. However, the OSNs medium allows legitimate and genuine users as well as spammers and other low trustworthy users to publish and spread their content leveraging the open environment and fewer restrictions (Abu-Salih 2018; Abu-Salih, Bremie, et al. 2019; Abu-Salih et al. 2020; Abu-Salih et al. 2018; Abu-Salih et al. 2019a, b; Chan et al. 2018; Meneghello et al. 2020; Wongthongtham and Salih 2018). Hence, it is vital to measure users’ credibility in numerous domains, thus indicating domain-based influential users, and filter out spammers and low trustworthy users.

This paper incorporates CredSaT (Abu-Salih et al. 2019a, b); a comprehensive credibility mechanism intended to measure users’ credibility based on their domains of knowledge. CredSaT provides an effective solution to discover spammers and influential domain-based users from the list of users whose domain(s) of knowledge is tacit, incorporating the temporal factor. The outcome of the credibility module is a ranked list of users with a corresponding credibility value for each specific domain. The temporal factor is assimilated in CredSat; the dataset of a user’s data

and metadata is divided into several chunks, where each chunk represents a specific period. A metric of credibility measurements is used to evaluate the user's trustworthiness in each particular chunk, thus providing overall credibility values. The mechanism used to calculate a user's value in each step considers other users' values, thereby providing a normalisation approach for building the relative ranking list of credibility in each domain. Hence, each particular key-value obtained from the user's data and metadata is measured against other users' values. In other words, each of the key attributes is normalised in each domain by dividing the value of the user's attribute by the maximum value achieved by all users in that domain. CredSaT shows the effectiveness of its embodied framework by benchmarking it against other state-of-the-art baseline models.

As mentioned previously our study uses the Twitter graph dataset crawled by Akcora et al. (2014). This dataset comprises spammers and other anomalous users. Hence, the main purpose of the knowledge credibility module is to filter out spammers and other low trustworthy users as their social contents affect the quality of the incorporated domain-based knowledge. For example, spammers who hijack tweets of politics-related contents, events, and stories should be eliminated from further conducted analysis despite the fact that political entities extracted from the contents of those users are relatively high. Table 1 shows the set of features incorporated into CredSaT framework. The reader can refer to (Abu-Salih et al. 2019a, b) to obtain further detailed explanations of the methodology used for measuring users' credibility.

As an example of the domain-based credibility analysis, Fig. 4 and Table 2 illustrate the key attributes used in the process of conducting credibility analysis on the social data and metadata collected for a well-known politician "Joanne Ryan@JoanneRyanLalor" as well as a social spammer "Ham—Hamjuku@hamjuku". Figure 4 illustrates the obtained values for certain domain-dependent attributes explained in Table 1. These values are computed for each of the 23 domains inferred from the domain discovery approach that is carried out utilizing IBM Watson API.

The values depicted in Fig. 4a demonstrate the domain-dependent analysis of Joanne's tweets which depicts a clear interest in the political domain of knowledge. This is evident considering that she is a member of the Australian House of Representatives and being active in this domain for several years.¹² Figure 4a also depicts that Joanne's tweets have had quite commended attention from her follower. This can be perceived due to the high number of domain-based likes, retweets, and replies. On the other hand, Fig. 4b shows the domain-based credibility analysis to a social spammer who demonstrated an interest in all domains. This commonly conveys a suspicious behaviour due to the following facts: (1) No one person is an expert in all domains (Gentner and Stevens 1983); (2) A user who posts in all domains does not convey to other users which domain(s) s/he is interested in. A user shows to other users which domain s/he is interested in by posting a wide range of contents in that particular domain; (3) There is the possibility that this user is a spammer due to the behaviour of spammers posting tweets about multiple topics (Wang 2010). This

¹² [https://en.wikipedia.org/wiki/Joanne_Ryan_\(politician\)](https://en.wikipedia.org/wiki/Joanne_Ryan_(politician)), accessed 24-03-2020.

Table 1 Selected features of CredSaT Framework

Feature	Description	Equation
Tweet Similarity Penalty (Twt_Sim)	Represents the count of unique keywords (#distinct-Words) in the overall user's tweets to the total number of keywords in the user's tweet (#words)	$Twt_Sim_u = \frac{\#DistinctWords_u}{\#Words_u}$
URL Similarity Penalty (URL_Sim)	Represents the percentage of non-redundant URLs (#DistinctURLs) with non-redundant hosts of URLs (#DistinctURLsHosts) to the total number of URLs (#URLs) posted by user u	$URL_Sim_u = 0.5 \times \left(\frac{\#DistinctURLs_u + \#DistinctURLsHosts_u}{\#URLs_u} \right)$
Domain-based content user score (Sum_cnt_scr)	Is computed by adding all scores retrieved from IBM Watson of tweets' texts posted by user u in domain d	–
Domain-based user URL scores (Sum_url_scr)	Is calculated by accumulating scores for all websites' content of the URLs embedded in user u 's tweets in domain d	–
Domain-based user scores (Sum_all_scr)	Refinement summing of the corresponding scores achieved by IBM Watson for all tweets' texts ($Sc_{u,d}^{Twt}$), and the refinement summing of scores retrieved from URLs' webpage content ($Sc_{u,d}^{URL}$) posted by a user u where a domain d . was inferred	$Sc_{u,d} = (Twt_Sim_u \times Sum_cnt_scr_{u,d}^{Twt} + URL_Sim_u \times Sum_url_scr_{u,d}^{URL})$
In frequency (DF)	Count of domains the user u has tweeted about	–
Inverse domain frequency (IDF)	Distinguishes users among the list of their domains of interest	$IDF_u = \log \left(\frac{n}{DF_u} \right)$
Weight (W)	Users weights in each domain	$W_{u,d} = Sc_{u,d} \times IDF_u$
Domain-based user's retweets (R)	Represents the frequency of retweets for user' content in each domain d	–
Domain-based user's likes (L)	Represents the percentage of likes/Favourites count for the users' content in each domain d	–
Main-based user's replies (P)	Embodies the count of replies to the users' content in each domain d	–

Table 1 (continued)

Feature	Description	Equation
Domain-based user positive sentiment replies (<i>SP</i>)	Refers to the sum of the positive scores of all replies to a user <i>u</i> , in domain <i>d</i> . Positive scores are achieved from IBM Watson with values greater than “0” and less than or equal to “1”. The higher the positive score, the greater is the positive attitude the repliers have to the users’ content	–
Domain-based user negative sentiment replies (<i>SN</i>)	Represents the sum of the negative scores of all replies to a user <i>u</i> in domain <i>d</i> . Negative scores are those values greater than or equal to “-1” and less than “0”. The lower the negative score, the greater is the negative attitude the repliers have to the users’ content	–
Domain-based user sentiments replies (<i>S</i>)	Embodies the difference between the positive and negative sentiments of all replies to user <i>u</i> , in the domain <i>d</i> .	$S_{u,d} = SP_{u,d} - SN_{u,d} $
Users’ followers (<i>FOL</i>)	Total count of users’ followers	–
’s friends <i>FRD</i>	Total count of user’s friends (followers)	–
Followers-friends ratio. <i>FF_R</i>	User followers-friends ratio	$FF_R_u = \begin{cases} \frac{FOL_u - FRD_u}{Age_u}, & fFOL_u - FRD_u \neq 0 \\ \frac{1}{Age_u}, & iFOL_u - FRD_u = 0 \end{cases}$, where <i>Age</i> is the age of user profile in years

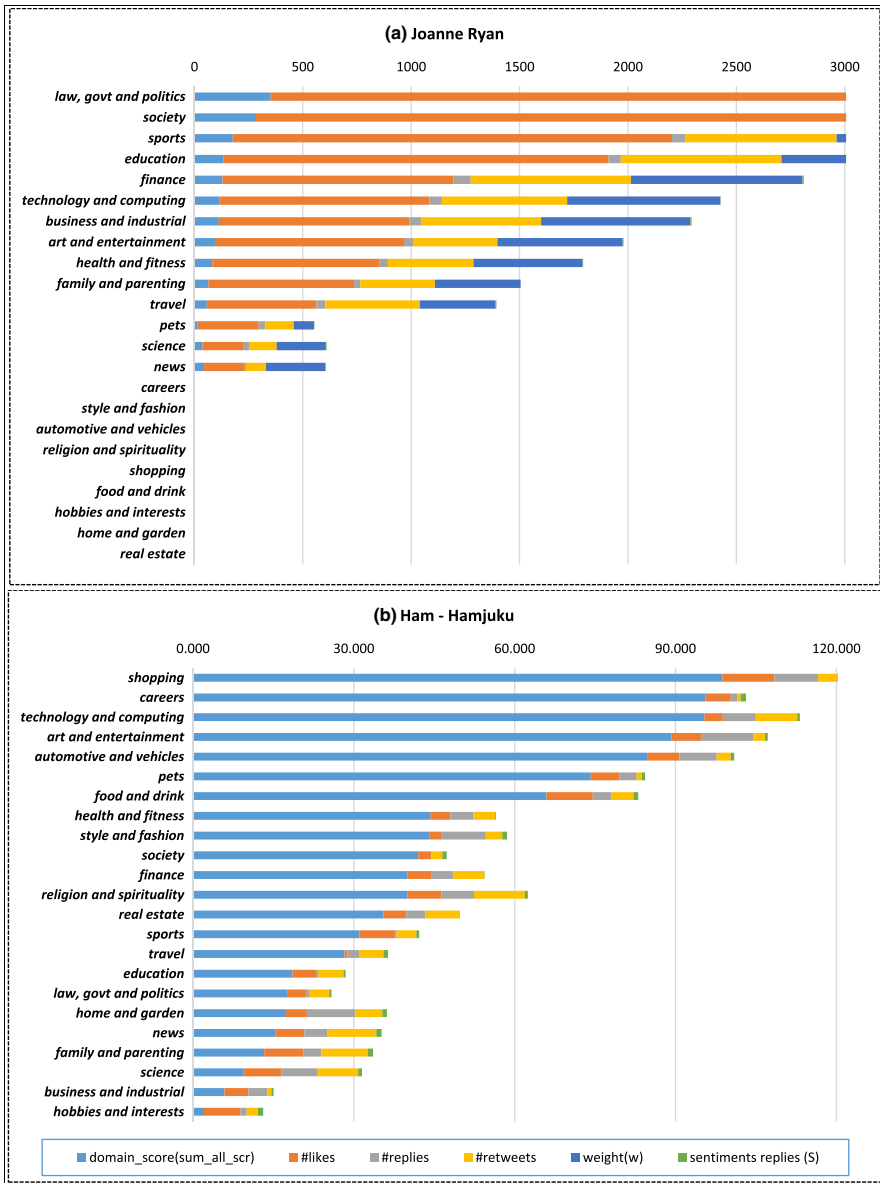


Fig. 4 Domain-dependent social data analysis of two twitterers: **a** Joanne Ryan (MPJoanneRyanLalor); a legitimate politician who is a member of the Australian house of representatives, **b** Ham—Hamjuku (hamjuku): a social twitterer spammer

could end up by tweets being posted in all domains which do not reflect a legitimate user’s behaviour as in the case of @hamjuku.

Further, Table 2 shows the domain-independent analysis to @JoanneRyanLalor and @hamjuku Twitter profiles. The figures exemplified in this table are plausibly

Table 2 Domain-independent social data analysis for a legitimate twitterer and a spammer twitterer

Feature	Joanne Ryan	Ham—Hamjuku
#followers	5606	248
#friends	1437	120
Age	7	13
TFFRatio	595.571	9.846
#Tweets	6459	3893
#DistinctWords	24,795	5392
#Words	112,889	10,733
Twt_Sim	0.212	0.502
#DistinctURLsHosts	85	5
#DistinctURLs	291	150
#URLs	861	5591
URL_Sim	0.218	0.528

acceptable; the number of users following *@JoanneRyanLalor*'s tweets is four times the total number of her friends (i.e., users who follow her). Also, Tweets and URL similarities computed for her 6495 tweets are around 20% which is quite reasonable. Differently, the similarity analysis computed for both tweets and URLs of *@hamjuku* poses a question on the quality of posted contents; publishing the same content repeatedly is obviously a spammer behaviour (Sedhai and Sun 2015). More than 50% of the tweets posted by *@hamjuku* are mainly repeated content. This implies to the textual contents as well as the embedded URLs. The TFF ratio that is calculated for *@hamjuku* sounds rational and legitimate considering the fact that the increase in the number of friends that a user u follows compared to the steadiness in the number of followers commonly indicates a suspicious behaviour, and such a user is likely to be a spammer (Twitter 2009; Wang 2010). However, as it can be inferred from the analysis conducted to *@hamjuku*, friends to followers ratio analysis could not be considered as sole spamming detection criterion, and this does not necessarily exhibit a credible profile; further scrutiny should be carried out to examine the overarching behaviour of a spammer, thereby providing a reliable detection mechanism.

3.5 Knowledge graph creation

At this stage, the knowledge representing the politics domain and the incorporated credible users and their data and metadata are captured in the domain ontology. In addition, knowledge is depicted in a less expressive relational model that stores knowledge obtained from the analysis conducted on users' social metadata and inferred from their collected textual content. The relational model embodies also the users' domain-based credibility indicating the trustworthiness of the users in each domain of knowledge. The knowledge graph creation module aims to transform the collected heterogeneous data format into a unified standard form.

The Resource Description Framework (RDF) is a widely used underlying model to represent knowledge in terms of triples (*subject, predicate, object*), where the

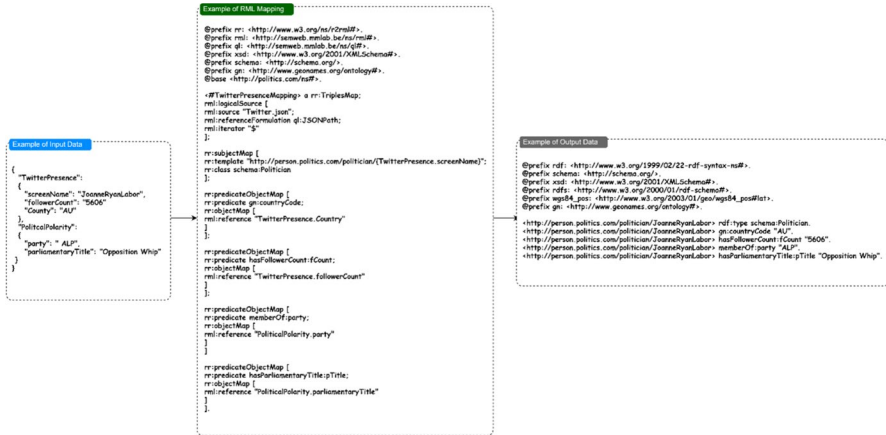


Fig. 5 Example of mapping a JSON data source to an RDF using RML

subject of the triple indicates the resource which needs to be described, predicate indicates the property of the subject, and object refers to the property value which describes the subject. A typical knowledge graph is represented as a directed graph where nodes indicate the entities (resources) of the class model and edges depicts the relations (properties) between those entities.

The datasets collected in this study are in different formats; Tabular, JSON, and CSV. One of the crucial steps in conducting Big data analytics is to provide a consolidated platform to handle the heterogeneity of the datasets collected from diverse data islands. Hence, we incorporate The RDF Mapping Language (RML) (Dimou et al. 2014) as a mapping language to express data in dissimilar format into a unified RDF form, thereby mitigating the variety dimension of Big data (Vidal et al. 2019). RML defines a generic approach for mapping different data structures, where the input could be any data source and the provided output is provided as an RDF graph. The mapping process in RML consists of one or more *triple maps*. In RML, each triple map embodies a logical source (input source), subject map (describes the mechanism to generate the subject for each logical resource), and predicate-object-map (specifies the predicate and the object map and how the triple's predicate is generated). RML mapping rules are used in this study to transform the annotated components into RDF triples to enrich the knowledge graph of the semantic repository. Figure 5 demonstrates an example of mapping an input JSON data source to RDF triples for an Australian politician (Joanne Ryan) using RML.

For annotation and enrichment, the domain knowledge graph is fed with annotated politics entities extracted from the textual contents of the tweets. The annotation is then enriched with a description of the concepts referring to the domain ontologies and using controlled vocabularies e.g., Dublin Core (DC¹³), Simple

¹³ <https://www.dublincore.org/>.

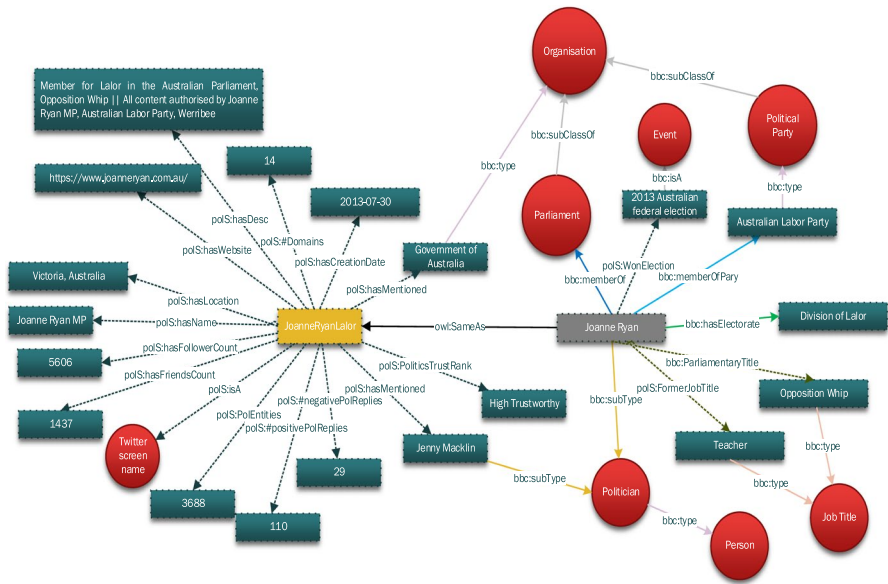


Fig. 6 Example of an RDF molecule describing “Joanne Ryan” obtained from the KG Creation process

Knowledge Organization System (SKOS¹⁴), Semantically-Interlinked Online Communities (SIOC¹⁵). This allows each entity in the textual data to be specified with its semantic concept. The particular concepts can be further expanded into other related concepts and other entities instantiated by the concepts. The consolidation of this semantic information provides a detailed view of the entities captured in domain ontologies.

For the interlinking process, entities are interlinked with similar entities defined in other datasets to provide an extended view of the entities represented by the concepts. Our focus is on equivalence links specifying URIs (Universal Resource Identifiers) that refer to the same resource or entity. Ontology Web Language (OWL) provides support for equivalence links between ontology components and 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. Hence, the data can be explored in further detail. In the interlinking process, different vocabularies i.e. Upper Mapping and Binding Exchange Layer (UMBEL), Freebase—a community-curated database of well-known people, places, and things, YAGO—a high-quality knowledge base, Friend-of-a-Friend (FOAF), Dublin Core (DC), Simple Knowledge Organization System (SKOS), Semantically-Interlinked Online Communities (SIOC), and Google KG, are used to link and enrich the semantic description of resources annotated.

¹⁴ <http://www.w3.org/2004/02/skos/>.

¹⁵ <http://sioc-project.org/>.

The domain KG is also enriched with knowledge inferred from the social presence of users on Twitter platform. This primarily encompasses associated metadata of the users and their content, such as #followers, #friends, #likes/favourites, and #retweet/share, etc. It also includes the resultants values obtained from the conducted domain-based credibility analysis to users. This includes values of the number of domains the users are interested in, the credibility value of the user in each domain of knowledge, the number of political entities indicated in the user's tweet, the number of the positive, negative and neutral replies to the user's tweets, etc.

Figure 6 illustrates an example of an RDF graph of knowledge inferred from multi-resources heterogeneous data collected for the politician, Joanne Ryan. This RDF graph can be referred to an RDF molecule as it represents a set of RDF triples indicating the same subject. This RDF molecule has been constructed as a result of the transformation process conducted on the data by the means of defined rules of RML mapping. The RML mapping rules are used further to ensure the format of the designated unique identifiers (URIs) for the mapped resources which are used as the subject of all the RDF triples.

3.6 Knowledge graph embedding models

Knowledge Graph Embedding (KGE) is the process of transforming the constituents of a KG (entities and relationships) into a low-dimension semantically-continuous space (Wang et al. 2017a, b). Even though solving problems pertaining to graphs can be carried out on the conventional graph presentation (i.e. adjacency matrix), mapping the entire graph or its nodes to the vector space has attracted the scientific community due to its scalability to simplify resolving several complex real-life graph problems such as KG completion, entity resolution, and link-based clustering, just to cite a few (Kipf and Welling 2016; Wang, Cui, et al. 2017; Nickel et al. 2015). Embedding a KG is learned via training a neural architecture over a graph, and comprises commonly three main components, namely; (1) encoding entities into distributed points in the vector space, and encoding relations as vectors, or other forms; (2) scoring function or model-specific function that is used to evaluate the model's efficiency; (3) optimization procedure, which aims to learn the optimal embedding for the designated KG, thereby the scoring function assigns high scores to positive statements.

The literature in KG Embedding commonly categorises the embedding techniques into two main classes; translation distance models and semantic matching models (Wang et al. 2017a, b). Translation Distance Models are designed to evaluate the plausibility of a certain fact in a distance between two entities. Semantic Matching Models intends to measure the plausibility of facts considering the latent semantics of entities and relations into their low dimensional representations. Amongst numerous KG embedding models proposed in the literature, the following are the set of most popular KG embedding models that are incorporated in this study.

Translating embedding (TransE) (Bordes et al. 2013) learns the representation of both the entities and relations as vectors in the same low dimensional semantic space. Hence, for a golden triple (h, r, t) , **TransE** treats the relation r as a translation

in the embedding space so that $h + r \approx t$, when (h, r, t) holds (t should be the closest to $h + r$), otherwise $h + r$ should be away in distance from t .

The *DistMult* model (Yang et al. 2014) is an extension and a simplification to RESCAL (Nickel et al. 2011) and is based on the bilinear model. In this model, the relation is encoded as diagonal (single vector) using the trilinear dot product as a scoring function.

Complex embeddings (ComplEx) (Trouillon et al. 2016) this is an extension to DistMult model by introducing complex-valued embeddings, where the scoring function is based on the trilinear Hermitian dot product in C . Entity and relation embeddings are no longer positioned in real space but in a complex space.

Holographic embeddings (Hole) (Nickel et al. 2016) a compositional vector space model that learns compositional vector space representations of entities and relations through incorporating the strength of RESCAL as well as the simplicity of **DistMult**.

Convolutional 2D KG embeddings (ConvE) (Dettmers et al. 2018) is a neural link prediction model that uses deep, multi-layer, conventional and fully connected layers of nonlinear features to tackle the interactions between input entities and relations.

Convolution-based model (ConvKB) (Nguyen et al. 2017) incorporates conventional neural networks to represent the concatenation of entities and relations, which increases the learning ability of latent features.

4 Experimental results

4.1 Dataset selection

As indicated previously this study aims to construct a domain-based KG (politics) and to carry out embeddings on the constructed KG that will assist in conducting further analysis. We make use of the Twitter platform to consolidate the domain-based KG with facts inferred from social contents propagated from this virtual platform. As indicated in Sect. (3.2.1 Dataset acquisition), the dataset is collected from dissimilar resources based on three different categories of users: (A) Members of the Australian house of representative (Senators and MPs); (B) users interested in Politics domain; and (C) users whose domain of interest is not explicitly conveyed. This set also might contain spammers, anomalous users, and other untrustworthy users. Domain analysis using IBM Watson has been conducted on each category of users to infer the domains of interest for each category. Figure 7 illustrates the total number of users and their tweets distributed over 23 domains of knowledge for each designated category.

As depicted in Fig. 7, category (A) shows a clear interest in the political domain which is reasonable (i.e. users of this category are mainly politicians, and their social contents are expected to discuss topics related to politics). Category (B) is a mixture of users who are selected as they explicitly show a common interest in politics. The domain analysis on category (B) supports this and shows that those users are interested in politics as well as other domains such as technology,

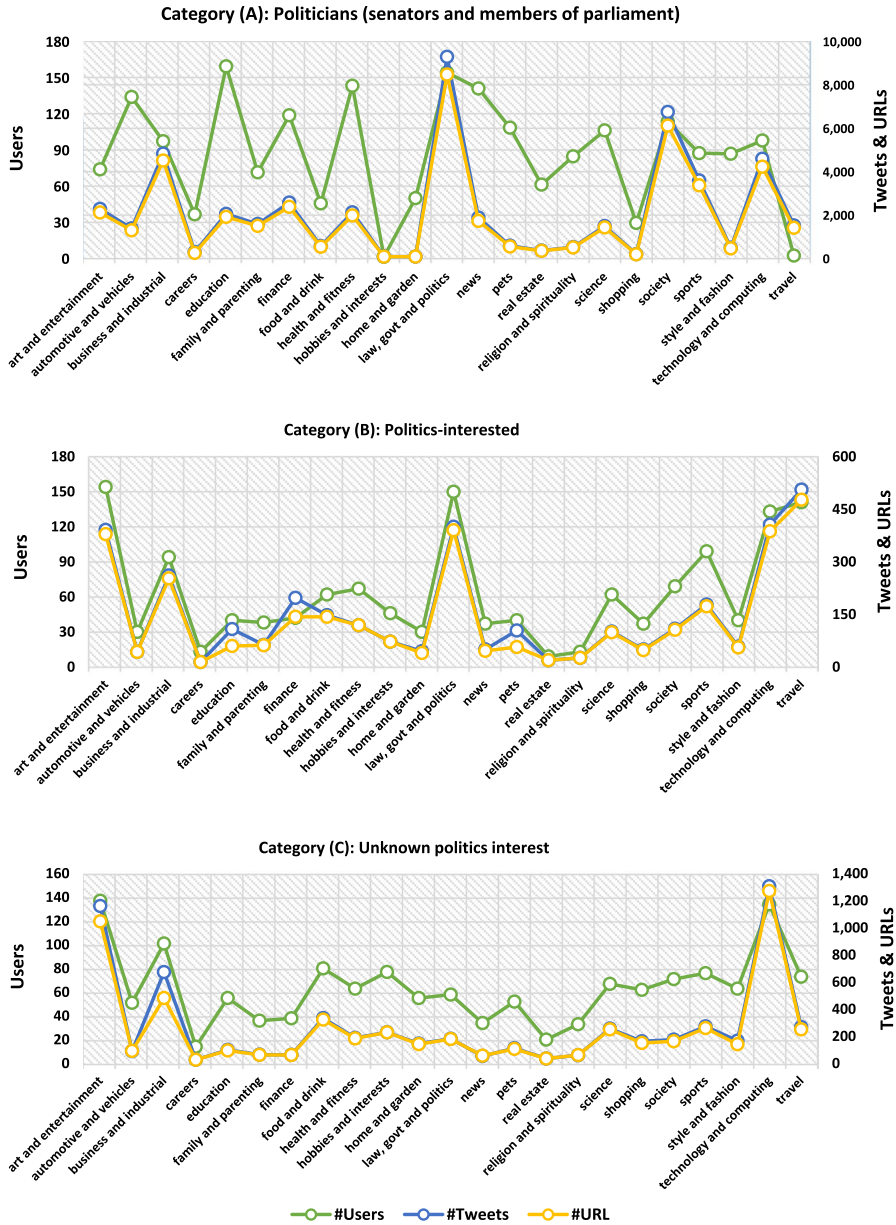


Fig. 7 The distribution of the total number of users and posted tweets and URLs in each designated domain for three categories: a Politicians (senators and members of parliament), b Politics-interested and c Unknown politics interest

art and entertainment, and travel. Despite the slight interest in politics domain and a strong interest in other venues, Category (C) demonstrates a balanced interest across the topics of interest.

Table 3 Datasets before and after conduction social credibility analysis

Users category	#users		#tweets		#entities		#political_entities		#facts(triples)	
	Before	After	Before	After	Before	After	Before	After	Before	After
(A) Politicians (senators and members of parliament)	227	227	611,739	611,739	1,461,645	1,461,645	172,095	172,095	16,286	16,286
<i>Low-trustworthy</i>	0		0		0		0		0	
(B) Politics-interested	622	560	1,540,723	1,386,651	1,815,091	1,542,827	136,458	118,718	34,149	27,319
<i>Low-trustworthy</i>	10%		9%		15%		13%		20%	
(C) Unknown politics interest	253	114	221,840	110,920	845,293	338,117	6,633	2,454	3,062	1,071
<i>Low-trustworthy</i>	55% (15% spam)		50%		60%		63%		65%	

Table 4 Hyperparameters used for the KG Embedding models

Hyperparameter	Description
Batches_count	<i>#batches to complete one epoch of the Platt scaling training</i>
Seed	<i>The value of the seed used to infer random numbers</i>
Epochs	<i>#iterations used to train the Platt scaling model</i>
k	<i>Embedding space dimensionality</i>
Eta	<i>#Negatives that must be generated at runtime during training for each positive</i>
Loss	<i>Loss function used to train the model. Examples: { pairwise, nll, absolute_margin, self_adversarial, multiclass_nll }</i>
Loss_params	<i>Set of parameters used for loss-specific hyperparameters</i>
Regularizer	<i>The strategy used with the loss function</i>
Regularizer_params	<i>Set of parameters for regularizer-specific hyperparameters</i>
Optimizer	<i>The optimizer used to update the weight parameters thereby minimizing the loss function. Examples: { 'sgd', 'adagrad', 'adam', 'momentum' }</i>
Optimizer_params	<i>Arguments passed as parameters to the optimizer, such as learning rate(lr) and momentum</i>
Verbosebool	<i>The verbose mode</i>

Domain-based social credibility module aims to carry out scrutiny on the collected and pre-processed dataset before conducting further analysis, thereby augmenting the KG with facts obtained from users who are legitimate and convey credible interest in politics domain, and it also aims to eliminate contents of users who obtained low trustworthiness values as per the mechanism discussed in Sect. 3.4.

The social credibility module was applied on the dataset and thus generated a new dataset embodying legitimate and politics-interested users with their associated contents. Table 3 shows some figures on the collected datasets for each category before and after conducting the social credibility module.

As demonstrated in Table 3, the contents belong to Category (A) was kept as it is with no cleansing. This is because this dataset comprises selected users (senators and members of the Australian House of Representatives). These users are politician and their social content involves the main source for KG creation. The percentage values of low-trustworthy users, tweets, entities, political_entities, and facts for the remaining categories are justifiable. Category (B) contains the users who might show a certain interest in the political domain but not necessarily a genuine interest. Therefore, ten percent (10%) of users in this category obtain low-trustworthy values. Category (C) contains a subset of users indicated in the Twitter graph dataset collected by (Akcora et al. 2014). As discussed in the Dataset Acquisition section, this dataset was used in the literature to discover spammers and other illegitimate accounts. Therefore, it is anticipated that the percentage value of low-trustworthy users in this category is higher than the former ones. In fact, fifty-five percent (55%) of users in this category are detected as low-trustworthy users (15% are spammers).

Table 5 The Embedding model and the incorporated hyperparameters used in the random search

Embedding model	Examined hyperparameters (optimal settings obtained from the random search are in bold format)
TransE	<i>batches_count</i> = {50, 100 , 150}, <i>seed</i> = {0, 555 }, <i>epochs</i> = {500, 1000, 2000 , 4000}, <i>k</i> = { 100 , 200}, <i>eta</i> = {5, 10, 15, 20}, <i>optimizer</i> = {'adam', ' adagrad '}, <i>loss</i> = {'pairwise', 'nll', 'absolute_margin'}, <i>verbose</i> = {True, False }, <i>regularizer</i> = {None, ' LP '}
DistMult	<i>batches_count</i> = {50, 100, 150}, <i>seed</i> = {0, 555}, <i>epochs</i> = {500, 1000, 4000 }, <i>k</i> = { 100 , 200}, <i>eta</i> = {5, 10, 15 , 20}, <i>optimizer</i> = {'adam', ' adagrad '}, <i>loss</i> = {'pairwise', 'nll', 'absolute_margin'}, <i>verbose</i> = {True, False }, <i>regularizer</i> = {None, ' LP '}, <i>normalize_ent_emb</i> = { True , False}
ComplEx	<i>batches_count</i> = {50, 100, 150, 200 }, <i>seed</i> = {0, 555}, <i>epochs</i> = {500, 1000 , 4000}, <i>k</i> = {100, 200 }, <i>eta</i> = {5, 10, 15 , 20}, <i>optimizer</i> = {'adam', ' adagrad '}, <i>loss</i> = {'pairwise', 'nll', 'absolute_margin'}, <i>verbose</i> = {True, False }, <i>regularizer</i> = {None, ' LP '}
HolE	<i>batches_coun</i> = {50, 100 , 150}, <i>seed</i> = {0, 555}, <i>epochs</i> = {500, 1000, 4000 }, <i>k</i> = {100, 200 }, <i>eta</i> = {5, 10, 15, 20}, <i>optimizer</i> = {'adam', ' adagrad '}, <i>loss</i> = {'pairwise', ' nll ', 'absolute_margin'}, <i>verbose</i> = { True , False}, <i>regularizer</i> = {None, ' LP '}
ConvE	<i>batches_count</i> = {50, 100 , 150}, <i>seed</i> = {0, 555 }, <i>epochs</i> = {500, 1000 , 4000}, <i>k</i> = {100, 200}, <i>eta</i> = {5, 10, 15 , 20}, <i>optimizer</i> = {'adam', 'adagrad'}, <i>loss</i> = {' BCE '}, <i>verbose</i> = {True, False }, <i>regularizer</i> = {None, ' LP '}, <i>conv_filters</i> = {24, 32 }, <i>conv_kernel_size</i> = {1, 2, 3}, <i>dropout_embed</i> = {0.2, 0.3}, <i>dropout_conv</i> = {0.2, 0.3}, <i>dropout_dense</i> = {0.1, 0.2 }
ConvKB	<i>batches_count</i> = {50, 100, 150 }, <i>seed</i> = {0, 555 }, <i>epochs</i> = {500, 1000, 4000 }, <i>k</i> = {100, 200 }, <i>eta</i> = {5, 10 , 15, 20}, <i>optimizer</i> = {'adam', ' adagrad '}, <i>loss</i> = {'pairwise', 'nll', 'absolute_margin'}, <i>verbose</i> = { True , False}, <i>regularizer</i> = {None, ' LP '}, <i>num_filters</i> = {24, 32}, <i>filter_sizes</i> = {1, 2, 3}, <i>dropout</i> = {0.0, 0.1 }

4.2 Domain KG embedding model evaluation

4.2.1 Experiment settings

This study incorporates AmpligraphTM (Costabello et al.) version 1.3.1, with TensorFlow 1.14, and Python 3.7 on the backend for conducting KG embeddings on the constructed domain KG. All the experiments including training and evaluation of each embedding model were carried out using the Australian Pawsey supercomputing high-performance facilities.¹⁶ The domain KG is initially divided into training, test, and validation subsets. Several KG embedding models are implemented and their hyperparameters are tuned using the random search strategy. Random search has proven efficiency and outperformed grid search routine as it provides a solid baseline, and it also shows robustness when the number of parameters increases (Bergstra and Bengio 2012; Li et al. 2016, 2017). A brief description of some internal settings used in the incorporated embedding models is provided in Table 4.

¹⁶ <https://pawsey.org.au/>.

Table 6 Comparison analysis of evaluation metrics of the embedding models using two KGs, where KG1 is the KG that is constructed based on the original dataset before applying credibility module, and KG2 is the constructed KG on the curated dataset after applying the credibility module

Embedding model	Hits@1		Hits@3		Hits@10		MRR	
	KG1	KG2	KG1	KG2	KG1	KG2	KG1	KG2
TransE	0.092	0.205	0.172	0.555	0.400	0.590	0.432	0.528
DistMult	0.129	0.215	0.354	0.515	0.522	0.641	0.389	0.580
ComplEx	0.129	0.465	0.362	0.554	0.124	0.640	0.416	0.622
HolE	0.123	0.480	0.189	0.660	0.756	0.790	0.305	0.679
ConvE	0.465	0.501	0.440	0.680	0.480	0.729	0.398	0.680
ConvKB	0.297	0.550	0.597	0.669	0.720	0.810	0.757	0.761

Table 5 shows the set of hyperparameters tested using the random search strategy, and those underlined are the optimal values obtained from the incorporated search strategy.

4.2.2 Evaluation protocol

This study incorporates the evaluation protocol proposed by Bordes et al. (2013). There are three key steps in the defined protocol, namely: (1) generating negative triples synthetically; (2) remove the resultant positive triplets; then (3) ranking all the test facts (triples) against the triples returned from the preceding step. Negative triples are initially positive triples (correct facts) which have been manipulated (corrupted) by randomly replacing head, tail or relation, thus creating new triples (false facts). The negative sampling mechanism used in this paper is based on corrupting the h Poggio2016, then we compute the average of attained evaluation metrics of each method.

To evaluate the effectiveness of the social credibility module, we carry out the aforementioned protocol on two different KGs, namely a *KG1* that is generated by accumulating the original dataset as demonstrated in Table 3 before applying the credibility module, and a *KG2* that is generated from the datasets on which the credibility module has been applied to (curated dataset). The next subsection provides a further discussion on the conducted experiments.

4.2.3 Embedding evaluation results

The experiments have been carried out incorporating six well-known embedding models as depicted in Table 4 along with the depicted tuned hyperparameters. With the ranks obtained from the subjects and predicates corruption of each dataset described in the previous subsection, the metrics are computed for each embedding model on each generated KG.

Table 6 illustrates the attained metric values obtained from each model on each KG. Two key findings can be inferred from the figures illustrated in Table 6. First,

Table 7 Performance Comparison on Link Prediction task for six KG embedding models on two KGs

Embedding model	Accuracy		Precision		Recall		F1 Score	
	KG1	KG2	KG1	KG2	KG1	KG2	KG1	KG2
TransE	0.326	0.475	0.427	0.508	0.237	0.423	0.305	0.462
DistMult	0.244	0.473	0.420	0.559	0.312	0.47	0.358	0.511
ComplEx	0.306	0.45	0.329	0.468	0.541	0.589	0.409	0.522
HolE	0.408	0.556	0.507	0.632	0.393	0.679	0.443	0.655
ConvE	0.601	0.604	0.647	0.72	0.469	0.689	0.544	0.704
ConvKB	0.616	0.744	0.597	0.832	0.407	0.722	0.484	0.773

all embedding models perform well on KG2 in comparison with KG1. This extends the significance of incorporating a credibility module for data purification, particularly data collected from mixed-quality resources such as social media. Second, despite the convergence in the outcome performance results, ConvKB embedding model outperforms other models in most metrics. For example, examining *hits@1*, *hits@3*, and *hits@10*, with ConvKB, we were able to hit a correct subject or predicate 55%, 66.9%, and 81% of the times respectively using KG2. This interpretation applies to all other metric values obtained from each embedding model. The good performance of ConvKB is established due to the underlying structure of ConvKB; it incorporates a CNN network to capture the global relationships and the transitional features of the KG embodied entities and relations. HolE model has also shown promising results; this is understandable as Hole integrates the efficiency and simplicity of more than one model (Wang et al. 2017a, b). Moreover, Hole can obtain rich interactions in such relational data by applying circular correlation on vectors that create compositional representations.

The utility of the embedding models is commonly measured by the applicability of using these models in more factual tasks. The following sections discuss the utility of the developed approach in link prediction, clustering, and visualisation tasks.

4.3 Experiments on downstream tasks

4.3.1 Task (1): link prediction

The implemented KG Embeddings in this study are used to carry out a Link Prediction task. We have generated a set of facts in politics domain, which contain true political facts that have not been trained in the model (unseen facts) as well as some synthetically created false politics facts. The goal is to test the utility of the model to detect which of the presented true candidate facts are likely to be true. Similarly, which false candidate facts are unlikely to be true. To evaluate the performance of this task, *accuracy*, *precision*, *recall*, and *F-measure* metrics are incorporated.

Precision specifies the proportion between the sum of actual true politics facts that are accurately predicted and the total sum of accurate and inaccurate predictions

of true politics facts. Recall specifies the proportion between the number of actual true politics facts that are accurately predicted and the total sum of actual true politics facts. Therefore, obtaining a high precision value indicates that the prediction module is a success in the result relevancy measure and is able to deduce more relevant politics facts among the retrieved ones. Attaining a high recall value indicates that the prediction module is a success in retrieving truly relevant results. For example, if a prediction module is evaluated and attains a precision value of ‘1’, this indeed conveys that all predicted facts are correct predictions and depict factual politics facts that can be used to augment the knowledge graph. However, this does not necessarily reflect the module’s efficacy to retrieve all true politics facts. On the other hand, if the prediction module attains a recall value of ‘1’, this implies that this module is able to retrieve all true positive facts. Yet, it does not convey the number of other false retrieved predictions. This is why it is commonly a good practice to incorporate the F-measure metric as it provides a weighted average of precision and recall.

The ground truth dataset of the link prediction experiment contains 1,000 labelled statements of both true politics facts as well as false politics facts obtained equally from two generated KGs (original and curated). Table 7 shows the performance comparison of the link prediction task for six incorporated embedding models and the two KGs. As depicted in the table, ConvKB embedding model has relatively overshadowed other embedding models in this task for both KGs. For example, this designated model has obtained 74.4%, 83.2%, 77.2%, 77.3% in accuracy, precision, recall, and f-score metrics respectively using KG2. HoIE and ConvE have also shown promising performance results. For example, ConvE was able to predict almost half of the true positive facts correctly as a true positive statement. Also, the precision computed for ConvE was 72% which proves the ability of this embedding model to demonstrate virtuous results in this task.

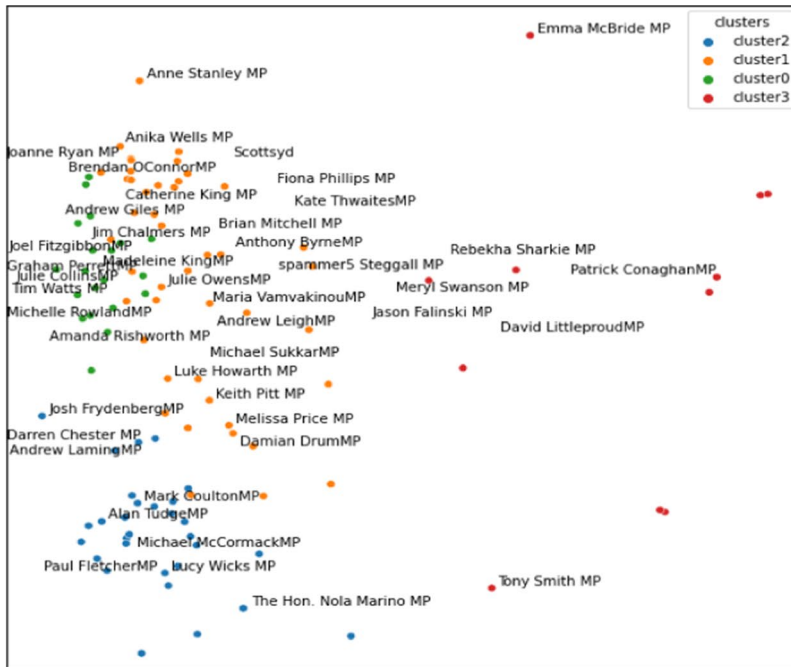
On the other hand, Table 7 shows that the results of TransE, DistMult and ComplEx embedding models are convergent in almost all computed metrics (i.e. accuracy, precision, recall, and f1-score) for both KGs. In spite of the fact that TransE performs well in datasets that embody one-to-one relationships, it demonstrates inadequacy to handle unbalanced relations (i.e. one-to-many/many-to-one) (Rossi et al. 2020). For example, embedding two knowledge facts such as; (Anthony Albanese, hasLocation, NSW) and (John Alexander, hasLocation, NSW) will result in “Anthony Albanese” entity vector be close to “John Alexander” entity vector. However, this does not convey the factual and realistic similarity between these two politicians; “Anthony Albanese” is a member of the Australian Labor Party while the affiliation of “John Alexander” is Liberal Party. This discrepancy also applies to their electorate and other facets. Furthermore, DistMult embedding model is inadequate to handle asymmetric and antisymmetric relations. This is evident because of the entry-wise product depicted in Eq. (2); it demonstrates that all relations are symmetric. This attains misleading results when asymmetric and antisymmetric relations are present (Wang, Ruffinelli, et al. 2018). HoIE, on the other hand, is skillful to address this issue since it uses a circular correlation operator, this results in HoIE able to capture the relations with asymmetry and anti-symmetry (Sharma and Talukdar 2018).

Table 8 A selected set of labelled candidates (true and false facts) from KG2

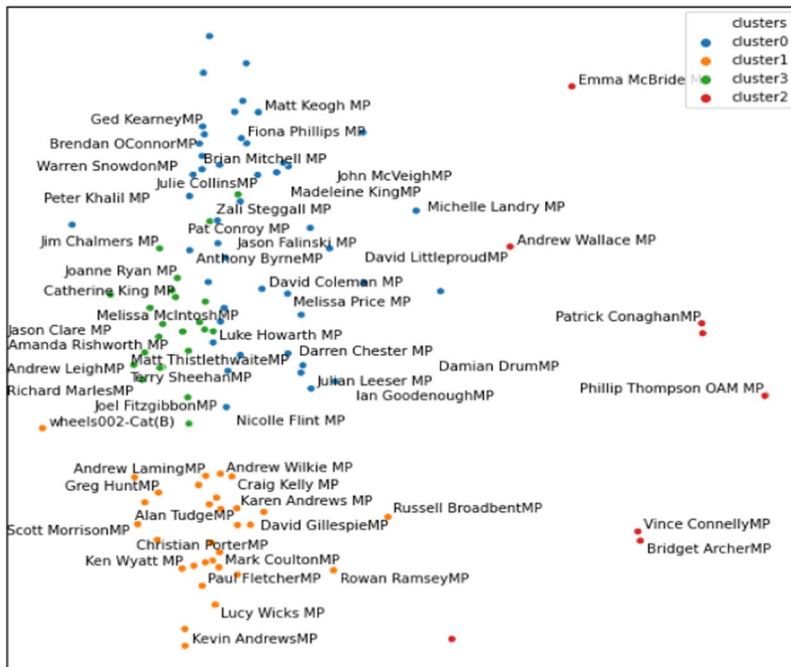
Unseen statement			Label	Prediction
Subject	Predicate	Object		
Karen Andrews MP	Member of party	Australian Labor Party	<i>True</i>	<i>True</i>
jocey70	Member of parliament	Australian Parliament	<i>False</i>	<i>True</i>
Nic Hodges	Has subtype	Politician	<i>False</i>	<i>True</i>
JohnKeily1	Has politics interest	High	<i>True</i>	<i>True</i>
Jay McCormack	Has mentioned	Australian Parliament	<i>False</i>	<i>False</i>
Karen Andrews MP	Member of parliament	Australian Parliament	<i>True</i>	<i>True</i>
Dgoodlad	Member of parliament	Australian Parliament	<i>False</i>	<i>True</i>
Bridget Archer MP	Has mentioned	Australian Labor Party	<i>False</i>	<i>False</i>
Bridget Archer MP	Member of party	Liberal Party of Australia	<i>True</i>	<i>False</i>
Kevin Andrews MP	Supports	Liberal Party of Australia	<i>True</i>	<i>True</i>
JohnKeily1	Supports	Australian Labor Party	<i>False</i>	<i>True</i>
Stevenc	Member of party	Australian Greens	<i>True</i>	<i>True</i>
Katie Allen MP	Member of party	Australian Labor Party	<i>False</i>	<i>False</i>

Further, Table 7 demonstrates the importance of incorporating the social credibility model. This can be seen in the relatively poor performance of all embedding models on link prediction task using the low-quality KG (i.e. KG1) and the better performance of the same embedding models on high quality and cleansed KG (i.e. KG2).

Table 8 shows an example of the link prediction task. The table presents a set of selected statements obtained from the ground truth of KG2, each statement with a label indicating, whether the statement is true or false along with the classification label acquired from ConvKB embedding model. It can be seen that the embedding model has been largely able to understand Australian politics and provide some good predictions on this domain. For example, the model is able to indicate that Karen Andrews is actually a member of the Australian Labor Party despite the fact that this information was not imported to the KG. It is also able to discover the political interest of users whose domain of interest is not explicitly depicted. For example, the collected tweets of @JohnKeily1 demonstrate that this user is interested in politics and does not support the Australian Labor Party (ALP). The model captures some truth about this user and did detect that this user is highly interested in politics, yet it fails to capture that @JohnKeily1 is not a supporter of ALP party. This can be understood considering that @JohnKeily1 has some negative tweets about ALP party and his presence in the vector space turns out to be close to those supporting this political party. This explanation can be also applied to other instances where the model was unable to correctly classify them. Hence, in future work, the KG will be further scrutinised and enhanced to embody for example the sentiments of the social contents, political polarisation, etc.



Clustering analysis of KG1 (original)



Clustering analysis of KG2 (curated)

Fig. 8 Clustering analysis of the constructed KGs

4.3.2 Task (2): KGE clustering and visualisation

Cluster analysis is another evaluation strategy that can be performed on a constructed KG. Clustering occurs on embedding space of both entities and relations and it is an effective evaluation strategy to measure the subjective quality of the KG embedding. The clustering projects the original embedding with the predetermined space size into a 2D space, then a subjective measure is carried out to evaluate the embeddings. Several clustering algorithms¹⁷ have been implemented and evaluated, such as *AffinityPropagation*, *AgglomerativeClustering*, *Birch*, *DBSCAN*, *FeatureAgglomeration* and *KMeans*. Several projections have been generated from these clustering modules, yet *KMeans* algorithm has proven effective due to the factual projections that are generated by this algorithm. This experiment follows the standard embedding space size (i.e. $k=100$) used by AmpliGraphTM.

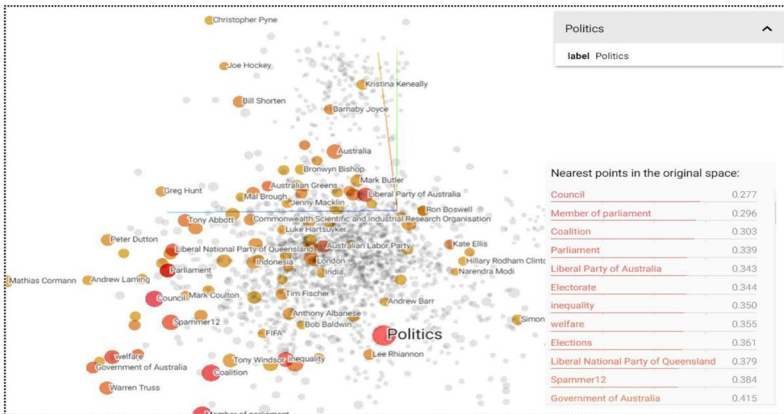
Figure 8 illustrates two figures, namely clustering analysis for KG1(original) and clustering analysis for KG2 (curated). Each analysis provides four clusters that show the semantic representation of entities(users) obtained from both datasets. The clustering analysis of KG1 that is constructed from the mixed-quality dataset (original) shows a social spammer (*spammer5-the twitter screen name is concealed*) that hijacks an orange-coloured cluster. The cohort of this cluster are mainly users who belong to or have an interest in the Austrian Labour Party. The clustering analysis of KG1 also demonstrates certain inadequacy in providing the correct assembly of legitimate users who have the same political belongings. For example, Melissa Price MP, a politician who belongs to the Liberal Party of Australia, is appended to the category of politicians who support or belong to the National Party of Australia.

On the other hand, Fig. 8—clustering analysis of KG2 shows a set of selected politicians and legitimate users interested in politics. The depicted cluster is widely accepted; it can be seen from the figure that almost all the members who have been categorised to the same group have the same political attachment. For example, Kevin Andrews MP, Lucy Wicks MP, Rowan Ramsey MP to name a few are all members of the Liberal Party of Australia and have been grouped to the same cluster. Likewise, Matt Keogh MP, Fiona Phillips, and Brian Mitchell are also assembled with others in the same cohort (i.e. ALP). Further, the figure depicts that the incorporated clustering approach is also able to infer politics affiliation of non-politicians; for example, the twitterer (@wheels002) happens to appear in the same vector space with members of the Australian Parliament who are also members of the Australian Greens Party. This is evident since @wheels002 has conveyed her interest in this political party in several tweets. Clustering analysis of these two KGs does verify the significance of incorporating social credibility module, not only to eliminate low-trustworthy social users and content but also to provide a better analysis on the downstream tasks.

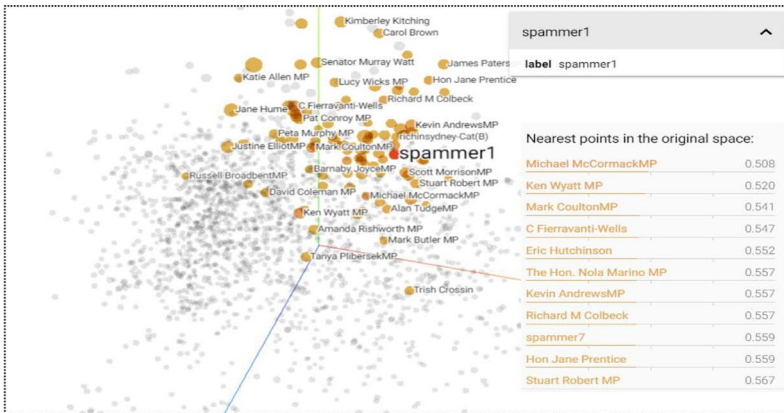
TensorBoard¹⁸ toolkit is used to visualise the implemented KG embeddings of the two KGs in 3D view and to project the resultant embeddings into low dimensional

¹⁷ <https://scikit-learn.org/stable/modules/clustering.html>.

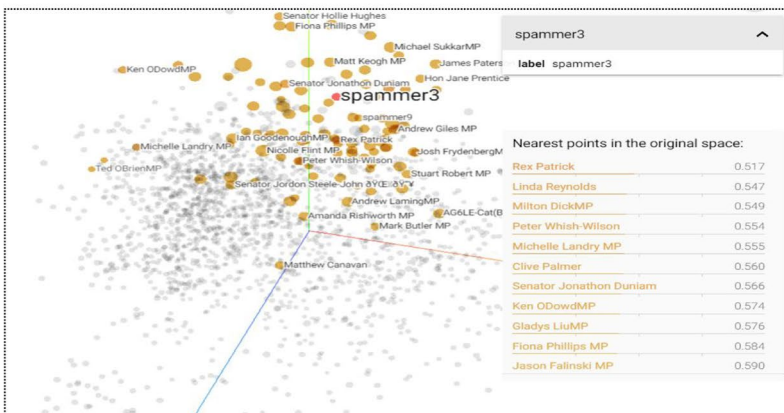
¹⁸ <https://www.tensorflow.org/tensorboard>.



Embedding Visualisation (KG1-A): shows concepts appear closely to "Politics" entity in same semantic vector space

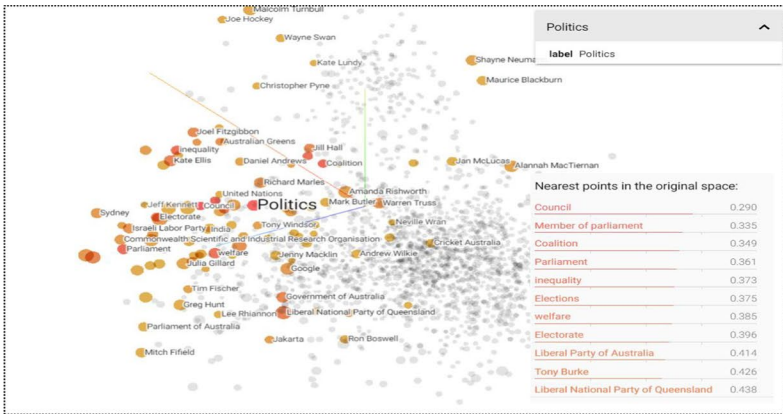


Embedding Visualisation (KG1-B): shows concepts appear closely to "spammer1" entity in same semantic vector space

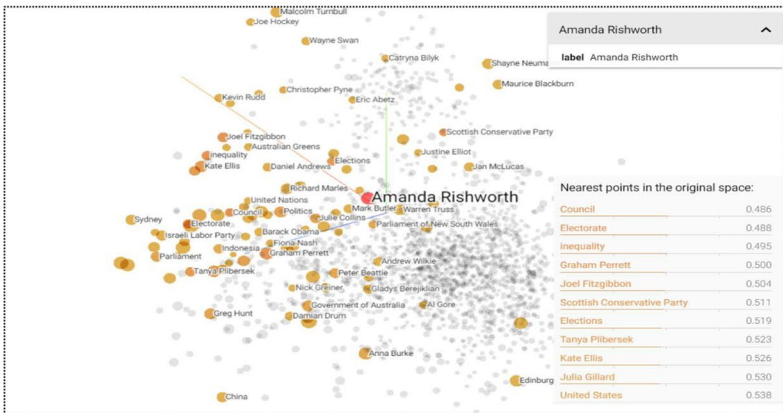


Embedding Visualisation (KG1-C): shows concepts that appear closely to "spammer3" entity in same semantic vector space

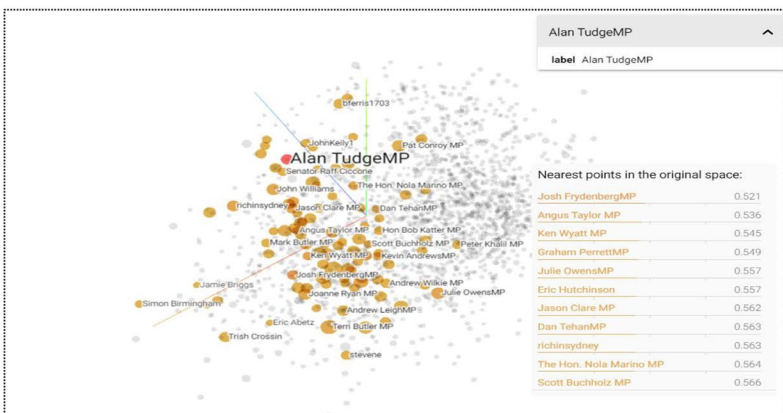
Fig. 9 KG Embedding visualisation using TensorBoard on KG1 (original)



Embedding Visualisation (KG2-A): shows concepts that appear closely to "Politics" entity in same semantic vector space



Embedding Visualisation (KG2-B): shows members of parliament who appear in the same semantic space with Amanda Rishworth MP



Embedding Visualisation (KG3-C): shows members of parliament who appear in the same semantic space with Alan Tudge MP

Fig. 10 KG Embedding visualisation using TensorBoard on KG2 (curated)

space using computed Principal Component Analysis (PCA). PCA is a statistical analyser that tends to minimise the dimensionality of a complex problem and explore patterns from the dataset by building a linear, multivariate model from the dataset (Rencher 2005). A dashboard of TensorBoard in PCA is used to provide an interesting 3D view of KG Embeddings. Figures 9a–c and 10a–c show visuals obtained from the TensorBoard on the original KG and the curated KG, respectively. These graphs illustrate the importance of visualisation to provide a subjective assessment of the implemented approach. Also, these visualisations demonstrate the significance of incorporating the social credibility module on KG embeddings. For example, Fig. 9a, listed all concepts related to “Politics” entity in the high-dimensional space. The nearest entities to Politics in the original space contains ‘spammers12’. This noisy data has been detected by the social credibility module, and thus eliminated from the dataset used to construct the curated KG (i.e. KG2). This is demonstrated in Fig. 10a in which the set of entities that are close to Politics are sound and convey the semantic relationships with the designated domain.

Figure 9b, c also illustrate the effect of applying KG embedding on KG that contains corrupted content (KG1). This is evident as spammer1 and spammer3 are positioned in the semantic space and thus convey relationships with legitimate politicians. On the other hand, Fig. 10b displays the cohort embedded with “Amanda Rishworth MP”, a member of ALP, using the curated KG. It is noticeable that the system is able to assemble members who factually share the common semantic features near to each other in the embedded space. This can also be supported by members appear with “Alan Tudge MP”, a member of LP, as depicted in Fig. 10c.

Using visualisation technique, as a subjective assessment to the implemented embedding approach, verifies again the applicability and utility of our approach on constructing high quality and trustworthy domain-based KG.

5 Conclusion and future work

The tremendous amount of information on the Web that is presented in dissimilar formats and covering various topics poses a challenge on possibilities to obtain the hoped-for added value from such massive data islands. This offers researchers a vital opportunity to consolidate efforts toward better understanding and analysis of such multimodal contents. In this context, Knowledge Graphs (KGs) are a popular phenomenon that has established a new venue by facilitating machines to understand meanings, thereby shrinking the semantic gap between them and humans. Further, domain-based KGs have extended these exerts by propagating knowledge in dissimilar domains that can be incorporated to resolve a variety of real-life problems. Yet, the credibility of knowledge is commonly neglected in the construction of KGs especially when the knowledge is harvested from social media where spammers and other low trustworthy users find a fertile medium to publish and spread their content taking advantage of the open environment and fewer restrictions of these platforms.

This paper proposes a credibility-based domain-specific KG Embedding framework. This framework incorporates certain modules which are integrated to manage and extract useful and trustworthy knowledge from the continuous propagation

of mixed quality social content. In particular, the framework is framed to contain the following modules: (1) Domain knowledge inference: Presents the core activity to prepare data for KG creation. It aims to detect and infer the user's domain of knowledge from pre-processed datasets. For a proof of concept purpose, the experiments are carried out on the Politics domain. The model makes use of various cross-domain knowledge-based repositories including Google KGTM, IBM Watson NLUTM, and WordnetTM to enrich the semantics of the textual contents, thereby facilitating the interoperability and integration to infer the political knowledge. (2) Social credibility module: A comprehensive credibility mechanism to measure users' credibility on politics domain incorporating a metric of key attribute. (3) KG construction: Aims to construct a politics KG leveraging politics ontology which captures knowledge representing the politics domain and the incorporated credible users and their data and metadata. (4) KG Embedding: this module incorporates state-of-the-art embedding models to embed the constructed KG in semantically interrelated and low dimensional vector space. Two KGs are used to demonstrate the utility of the constructed KG Embeddings, namely KG1 which is constructed using a poor and low-quality dataset, and KG2, curated KG, which is constructed using a cleansed version of the former dataset incorporating social credibility module. The embedding utility of these KGs is demonstrated and substantiated on link prediction, clustering, and visualisation tasks.

This paper is a report on work in progress as it is an ongoing project the purpose of which is to develop an integrated platform of various techniques for domain-discovery, credibility evaluation, and KG construction and embedding. Therefore, in the future work, the following extensions will be considered: (1) More embedding techniques will be implemented, and their evaluation will be examined. (2) CredSaT is the sole social credibility module that is used in this study, thus an array of other domain-based social credibility modules will be studied and implemented in order to consolidate the implemented approach. (3) Politics KG was constructed in this study for a proof of concept. In the future work, we will investigate other domains leveraging domain ontologies, semantic technologies and the linked open data cloud. (4) The current modules and new proposed enhancements will be automated and the entire architecture will be developed as an open-source project and will be released to facilitate replication and knowledge sharing.

References

- Abu-Salih B (2018) Trustworthiness in social big data incorporating semantic analysis, machine learning and distributed data processing. Curtin University
- Abu-Salih B (2021) Domain-specific Knowledge Graphs: a survey. *J Netw Comput Appl* 185:103076
- Abu-Salih B, Bremie B, Wongthongtham P, Kevin D, Tomayess I, Kit YC, Mohammad A, Teshreen A, Sulaiman A, Abdullah A, Muteeb A, Naser A, Abdulaziz A (2019) Social credibility incorporating semantic analysis and machine learning: a survey of the state-of-the-art and future research directions. In: 887–896. Springer International Publishing, Cham
- Abu-Salih B, Kit YC, Omar A-K, Marwan A-T, Wongthongtham P, Tomayess I, Heba S, Malak A-H, Bushra B, Abdulaziz A (2020) Time-aware domain-based social influence prediction. *J Big Data* 7:10


- Abu-Salih B, Wongthongtham P, Kit YC (2018) Twitter mining for ontology-based domain discovery incorporating machine learning *J Knowl Manag* 22:949–981
- Abu-Salih B, Wongthongtham P, Kit YC, Zhu D (2019) CredSaT: credibility ranking of users in big social data incorporating semantic analysis and temporal factor. *J Inf Sci* 45:259–280
- Akcora CG, Barbara C, Elena F, Murat K (2014) Detecting anomalies in social network data consumption. *Soc Netw Anal Min* 4:1–16
- Akrami F, Guo L, Wei H, Li C (2018) Re-evaluating embedding-based knowledge graph completion methods. In: *Proceedings of the 27th ACM international conference on information and knowledge management*, pp 1779–1782
- Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. *KDD* 2008 7:15
- Anderson M, Dennis Q (2020) 55% of U.S. social media users say they are ‘worn out’ by political posts and discussions. <https://pewrsr.ch/3aFYhtI>
- Balazevic I, Carl A, Timothy H (2019) Tucker: tensor factorization for knowledge graph completion. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp 5188–5197
- Balažević I, Carl A, Timothy MH (2019) Tucker: tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*
- Bastian M, Sebastien H, Mathieu J (2009) Gephi: an open source software for exploring and manipulating networks. In: *Third international AAAI conference on weblogs and social media*
- BBC (2014) BBC politics ontology. <http://www.bbc.co.uk/ontologies/politics>. Accessed 21 Sep 2019
- BBC Ontologies (2015). <http://www.bbc.co.uk/ontologies>. Accessed 19 May 2019
- Beheshti A, Boualem B, Reza N, Alireza T (2018) CoreKG: a knowledge lake service. *Proc VLDB Endow* 11:1942–1945
- Beheshti A, Boualem B, Quan ZS, Francesco S (2020) Intelligent knowledge lakes: the age of artificial intelligence and big data. In: *International conference on web information systems engineering*, pp 24–34. Springer
- Bergstra J, Yoshua B (2012) Random search for hyper-parameter optimization. *J Mach Learn Res* 13:281–305
- Bordes A, Nicolas U, Alberto G-D, Jason W, Oksana Y (2013) Translating embeddings for modeling multi-relational data. In: *Advances in neural information processing systems*, pp 2787–2795
- Cai H, Vincent WZ, Kevin C-CC (2018) A comprehensive survey of graph embedding: problems, techniques, and applications. *IEEE Trans Knowl Data Eng* 30:1616–1637
- Chan KY, Kwong CK, Wongthongtham P, Jiang H, Chris KYF, Bilal A-S, Liu Z, Wong TC, Pratima J (2018) Affective design using machine learning: a survey and its prospect of conjoining big data. *Int J Comput Integrated Manuf* 1–25
- Chen P, Yu L, Vincent WZ, Chen X, Boda Y (2018) Knowedu: a system to construct knowledge graph for education. *IEEE Access* 6:31553–31563
- Chen W, Xiao Z, Wang T, Bishan Y, Yi L (2017) Opinion-aware Knowledge Graph for Political Ideology Detection. In *IJCAI*, pp 3647–3653
- Costabello L, Sumit P, Chan LV, Rory M, Nicholas M, Pedro T (2019) Ampli-Graph: a library for representation learning on knowledge graphs, March 2019. 10.5281/zenodo.2595043
- Cui L, Haeseung S, Maryam T, Fenglong M, Suhang W, Dongwon L (2020) DETERRENT: knowledge guided graph attention network for detecting healthcare misinformation. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp 492–502
- Demchenko Y, Paola G, Cees DL, Peter M (2013) Addressing big data issues in scientific data infrastructure. In: *Collaboration Technologies and Systems (CTS), 2013 International conference on*, pp 48–55. IEEE
- Deng Y, Duo L, Dijiang H, Chun-Jen C, Fanjie L (2019) Knowledge graph based learning guidance for cybersecurity hands-on labs. In: *Proceedings of the ACM conference on global computing education*, pp 194–200
- Dettmers T, Pasquale M, Pontus S, Sebastian R (2018) Convolutional 2d knowledge graph embeddings. In: *Thirty-second AAAI conference on artificial intelligence*
- Dimou A, Miel VS, Pieter C, Ruben V, Erik M, Rik VDW (2014) RML: a generic language for integrated RDF mappings of heterogeneous data

- Dong X, Evgeniy G, Jeremy H, Wilko H, Ni L, Kevin M, Thomas S, Sun S, Wei Z (2014) Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 601–610
- Feng L (2020) Design of tourism intelligent recommendation model of Mount Tai Scenic area Based on Knowledge Graph. In: 2020 International conference on E-Commerce and Internet Technology (ECIT), pp 241–244 IEEE
- Getner D, Albert LS (1983) Mental models. L. Erlbaum Associates, Hillsdale
- Gong F, Meng W, Wang H, Sen W, Mengyue L (2021) SMR: Medical knowledge graph embedding for safe medicine recommendation. *Big Data Res* 23:100174
- Gonzalez JE, Reynold SX, Ankur D, Daniel C, Michael JF, Ion S (2014) "Graphx: graph processing in a distributed dataflow framework. In: 11th {USENIX} symposium on operating systems design and implementation ({OSDI} 14), pp 599–613
- Gruber TR (1993) A translation approach to portable ontology specifications. *Knowl Acquis* 5:199–220
- Gruhl D, Guha R, Liben-Nowell D, Tomkins A (2004) Information diffusion through blogspace. In: The 13th international world wide web conference (WWW'04), pp 491–501. ACM, New York, USA
- Halberstam Y, Brian K (2016) Homophily, group size, and the diffusion of political information in social networks: evidence from Twitter. *J Public Econ* 143: 73–88
- Han X, Liu Z, Sun M (2018) Neural knowledge acquisition via mutual attention between knowledge graph and text. In: Thirty-second AAAI conference on artificial intelligence
- Huang H, Larry H, Heng J (2015) Leveraging deep neural networks and knowledge graphs for entity disambiguation. arXiv preprint arXiv:1504.07678
- Huang L, Lin Z, Lv S, Lu F, Yue Z, Hu S (2017) KIEM: a knowledge graph based method to identify entity morphs. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 2111–2114
- Huang Z, Wei X, Kai Y (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991
- Ji S, Shirui P, Erik C, Pekka M, Philip SY (2020) A survey on knowledge graphs: representation, acquisition and applications. arXiv preprint arXiv:2002.00388
- Kazemi SM, David P (2018) Simple embedding for link prediction in knowledge graphs. In: Advances in neural information processing systems, pp 4284–4295
- Kejriwal M (2019) Domain-specific knowledge graph construction. Springer
- Kejriwal M, Runqi S, Pedro S (2019) Expert-guided entity extraction using expressive rules. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 1353–1356
- Kiesling E, Andreas E, Kabul K, Fajar E (2019) The SEPSES knowledge graph: an integrated resource for cybersecurity. In: International semantic web conference, pp198–214. Springer
- Kipf TN, Max W (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
- Kruit B, Peter B, Jacopo U (2019) Extracting novel facts from tables for knowledge graph completion. In: International semantic web conference, pp 364–381. Springer
- Laufer C, Daniel S (2017) On modeling political systems to support the trust process. In: PrivOn@ ISWC
- Li L, Kevin J, Giulia DS, Rostamizadeh AT, Hyperband A (2016) A novel bandit-based approach to hyperparameter optimization *Comput Vis Pattern Recognit*, arXiv: 1603.0656
- Li L, Kevin J, Giulia DS, Afshin R, Ameet T (2017) Hyperband: a novel bandit-based approach to hyperparameter optimization. *J Mach Learn Res* 18:6765–6816
- Li Y, Wei B, Liu Y, Liang Y, Hui C, Yu J, Zhu W (2018) Incorporating knowledge into neural network for text representation. *Expert Syst Appl* 96:103–114
- Liang X, Han C, Zhang W (2020) Knowledge extraction experiment based on tourism knowledge graph Q & A data set. In: 2020 IEEE International conference on power, intelligent computing and systems (ICPICS), pp 828–832. IEEE
- Liben-Nowell D, Kleiberg J (2008) Tracing information flow on a global scale using internet chain-letter data. *Natl Acad Sci* 105:4633–4638
- Lin D, Wu X (2009) Phrase clustering for discriminative learning. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, pp 1030–1038
- Lin Y, Liu Z, Sun M, Yang L, Xuan Z (2015) Learning entity and relation embeddings for knowledge graph completion. In: Twenty-ninth AAAI conference on artificial intelligence

- Liu H, Wu Y, Yiming Y (2017) Analogical inference for multi-relational embeddings. In: Proceedings of the 34th international conference on machine learning-Volume 70, pp 2168–2178. JMLR.org
- Liu J, Felix S, Keping L, Wei Z (2021) A knowledge graph-based approach for exploring railway operational accidents. *Reliab Eng Syst Saf* 207:107352
- Liu Y, Zeng Q, Joaquín OM, Yang H (2019) Anticipating stock market of the renowned companies: a knowledge graph approach. *Complexity*
- Long J, Chen Z, He W, Wu T, Ren J (2020) An integrated framework of deep learning and knowledge graph for prediction of stock price trend: an application in Chinese stock exchange market. *Appl Soft Comput* 106205
- Low Y, Joseph G, Aapo K, Danny B, Carlos G, Joseph MH (2012) Distributed graphlab: a framework for machine learning in the cloud. arXiv preprint arXiv:1204.6078
- Marin A, Roman H, Ruhi S, Mari O (2014) Learning phrase patterns for text classification using a knowledge graph and unlabeled data. In: Fifteenth annual conference of the international speech communication association
- Meilicke C, Manuel F, Wang Y, Daniel R, Rainer G, Heiner S (2018) Fine-grained evaluation of rule-and embedding-based systems for knowledge graph completion. In: International semantic web conference, pp 3–20. Springer
- Meneghelo J, Nik T, Kevin L, Kok WW, Bilal A-S (2020) Unlocking social media and user generated content as a data source for knowledge management. *Int J Knowl Manage (IJKM)* 16:101–122
- Morwal S, Nusrat J, Deepti C (2012) Named entity recognition using hidden Markov model (HMM). *Int J Nat Lang Comput (IJNLC)* 1:15–23
- Nakashole N, Raphael F (2017) Knowledge distillation for bilingual dictionary induction. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2497–2506
- Nguyen DQ, Tu DN, Dat QN, Dinh P (2017) A novel embedding model for knowledge base completion based on convolutional neural network. arXiv preprint arXiv:1712.02121
- Nguyen HL, Jason JJ (2019) Social event decomposition for constructing knowledge graph. *Futur Gener Comput Syst* 100:10–18
- Nickel M, Kevin M, Volker T, Evgeniy G (2015) A review of relational machine learning for knowledge graphs. *Proc IEEE* 104:11–33
- Nickel M, Lorenzo R, Tomaso P (2016) Holographic embeddings of knowledge graphs. In: Thirtieth Aaai conference on artificial intelligence
- Nickel M, Volker T, Hans-Peter K (2011) A three-way model for collective learning on multi-relational data. In: *Icml*, pp 809–816
- Palumbo E, Giuseppe R, Raphaël T (2017) Entity2rec: learning user-item relatedness from knowledge graphs for top-n item recommendation. In: Proceedings of the eleventh ACM conference on recommender systems, pp 32–36
- Pan JZ, Siyana P, Li C, Li N, Li Y, Liu J (2018) Content based fake news detection using knowledge graphs. In: International semantic web conference, pp 669–683. Springer
- Paulheim H (2017) Knowledge graph refinement: a survey of approaches and evaluation methods. *Semantic Web* 8:489–508
- Purohit H, Rajaraman K, Nikhil D (2019) Towards next generation knowledge graphs for disaster management. In: 2019 IEEE 13th international conference on semantic computing (ICSC), pp 474–477. IEEE
- Qiuyu D, Shang F (2020). Research on user knowledge acquisition and application in software ecology. In: *Journal of Physics: Conference Series*, 012030. IOP Publishing
- Rencher AC (2005) A review of “Methods of Multivariate Analysis”. In: Taylor & Francis
- Rossi A, Donatella F, Antonio M, Paolo M, Denilson B (2020) Knowledge graph embedding for link prediction: a comparative analysis. arXiv preprint arXiv:2002.00819
- Sedhai S, Aixin S (2015) Hspam14: a collection of 14 million tweets for hashtag-oriented spam research. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp 223–232
- Shapiro MA, Libby H (2017) Politicians and the policy agenda: Does use of Twitter by the US Congress Direct New York Times content? *Policy Internet* 9:109–132
- Sharma A, Partha T (2018) Towards understanding the geometry of knowledge graph embeddings. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 122–131
- Sheng M, Anqi L, Bu Y, Jing D, Yong Z, Xin L, Chao L, Xing C (2020) DSQA: a domain specific QA system for smart health based on knowledge graph. In: International conference on web information systems and applications, pp 215–222 Springer

- Shi B, Tim W (2018) Open-world knowledge graph completion. In: Thirty-second AAAI conference on artificial intelligence
- Song J (2019) Distilling knowledge from user information for document level sentiment classification. In: 2019 IEEE 35th international conference on data engineering workshops (ICDEW), pp 169–176. IEEE
- Stevens R (2001) What is an Ontology? Accessed 3rd March. <http://www.cs.man.ac.uk/~stevens/onto/node3.html>
- Sun E, Rosenn I, Marlow C, Lento T (2009) Gesundheit! modeling contagion through facebook news feed. In: ICWSM 2009. AAAI Press, San Jose, CA
- Tian F, Bin G, Qing C, Enhong C, Tie-Yan L (2014) Learning deep representations for graph clustering. In: Twenty-eighth AAAI conference on artificial intelligence
- Tong R, Xue L, Wang H (2016) Building and exploring an enterprise knowledge graph for investment analysis. In: Groth P, Simperl E, Gray A et al (eds) The semantic web-ISWC
- Trouillon T, Johannes W, Sebastian R, Éric G, Guillaume B (2016) Complex embeddings for simple link prediction. In: International conference on machine learning (ICML)
- Twitter (2009) The twitter rules. <https://support.twitter.com/articles/18311-the-twitter-rules>
- Van Kessel S, Remco C (2016) Shifting the blame. Populist politicians' use of Twitter as a tool of opposition. *J Contemp Eur Res* 12
- Vidal M-E, Kemele ME, Samaneh J, Farah K, Guillermo P (2019). Semantic data integration of big biomedical data for supporting personalised medicine. In: Current trends in semantic web technologies: theory and practice. Springer
- Wang AH (2010) Don't follow me: Spam detection in Twitter. In: Security and Cryptography (SECRYPT), Proceedings of the 2010 international conference on, pp 1–10
- Wang H, Zhang F, Wang J, Miao Z, Li W, Xing X, Guo M (2018) Ripplenet: propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th ACM international conference on information and knowledge management, pp 417–426
- Wang Q, Mao Z, Bin W, Li G (2017) Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans Knowl Data Eng* 29:2724–2743
- Wang X, Wang D, Canran X, He X, Cao Y, at-Seng C (2019) Explainable reasoning over knowledge graphs for recommendation. In: Proceedings of the AAAI conference on artificial intelligence, pp 5329–5336
- Wang X, Cui P, Jing W, Jian P, Zhu W, Yang S (2017) Community preserving network embedding. In: Thirty-first AAAI conference on artificial intelligence
- Wang Y, Daniel R, Rainer G, Samuel B, Christian M (2018) On evaluating embedding models for knowledge base completion. arXiv preprint arXiv:1810.07180
- West R, Evgeniy G, Kevin M, Sun S, Rahul G, Dekang L (2014) Knowledge base completion via search-based question answering. In: Proceedings of the 23rd international conference on World wide web, pp 515–526
- Wongthongtham P, Bilal AS (2018) Ontology-based approach for identifying the credibility domain in social Big Data. *J Organ Comput Electron Commer* 28:354–377
- Wu J, Zhu X, Zhang C, Zheng H (2020) Event-centric tourism knowledge graph—a case study of Hainan. In: International conference on knowledge science, engineering and management, pp 3–15. Springer
- Yang B, Wen-tau Y, He X, Gao J, Li D (2014) Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575
- Yang Z, William WC, Ruslan S (2016) Revisiting semi-supervised learning with graph embeddings. arXiv preprint arXiv:1603.08861
- Yao L, Yin Z, Wei B, Zhe J, Rui Z, Zhang Y, Chen Q (2017) Incorporating knowledge graph embeddings into topic modeling. In: Thirty-first AAAI conference on artificial intelligence
- Zhang Y, Dai H, Zornitsa K, Alexander JS, Le S (2018) Variational reasoning for question answering with knowledge graph. In: Thirty-second AAAI conference on artificial intelligence
- Zheng Y, Ruifang L, Hou J (2017) The construction of high educational knowledge graph based on MOOC. In: 2017 IEEE 2nd information technology, networking, electronic and automation control conference (ITNEC), pp 260–63. IEEE

Authors and Affiliations

Bilal Abu-Salih¹  · **Marwan Al-Tawil**¹ · **Ibrahim Aljarah**¹ · **Hossam Faris**^{1,2} · **Pornpit Wongthongtham**³ · **Kit Yan Chan**⁴ · **Amin Beheshti**⁵

✉ Bilal Abu-Salih
b.abusalih@ju.edu.jo

¹ King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan

² School of Computing and Informatics, Al Hussein Technical University, Amman, Jordan

³ The University of Western Australia, Perth, Australia

⁴ Curtin University, Perth, Australia

⁵ Macquarie University, Melbourne, Australia