Kingston *et al. Hum. Cent. Comput. Inf. Sci.* (2018) 8:22
https://doi.org/10.1186/s13673-018-0145-6

Human-centric Computing and Information Sciences

RESEARCH Open Access

Using semantic clustering to support situation awareness on Twitter: the case of world views

Charlie Kingston[1], Jason R. C. Nurse[2*], Ioannis Agrafiotis[1] and Andrew Burke Milich[3]

*Correspondence:
j.r.c.nurse@kent.ac.uk
[2] School of Computing, University of Kent, Canterbury, UK
Full list of author information is available at the end of the article

Abstract

In recent years, situation awareness has been recognised as a critical part of effective decision making, in particular for crisis management. One way to extract value and allow for better situation awareness is to develop a system capable of analysing a data-set of multiple posts, and clustering consistent posts into different views or stories (or, 'world views'). However, this can be challenging as it requires an *understanding* of the data, including determining what is consistent data, and what data corroborates other data. Attempting to address these problems, this article proposes *Subject-Verb-Object Semantic Suffix Tree Clustering* (SVOSSTC) and a system to support it, with a special focus on Twitter content. The novelty and value of SVOSSTC is its emphasis on utilising the Subject–Verb–Object typology in order to construct semantically consistent world views, in which individuals—particularly those involved in crisis response—might achieve an enhanced picture of a situation from social media data. To evaluate our system and its ability to provide enhanced situation awareness, we tested it against existing approaches, including human data analysis, using a variety of real-world scenarios. The results indicated a noteworthy degree of evidence (e.g., in cluster granularity and meaningfulness) to affirm the suitability and rigour of our approach. Moreover, these results highlight this article's proposals as innovative and practical system contributions to the research field.

Keywords: Social media analytics, User-support tools, Data clustering, Information systems, Crisis response, Computational social science, Fake news

Introduction

The quantity of information available online is astounding. Social media has played a key part in this boom of content, as it has emerged as a central platform for communication and information sharing, allowing users to post messages related to any event or topic of interest [1]. Possibly one of the most significant uses of social media today is its ability to help understand on-going situations. Situation awareness has been recognised as a critical part of effective decision making, in particular for crisis management scenarios [2]. Twitter, for instance, is regularly used as a news breaking mechanism to provide near-real-time observations about situations [3]. By leveraging the public's collective intelligence, emergency responders may be able create a holistic view of a situation, allowing them to make the most informed decisions possible.

There are two key problems that prevent users from gathering valuable and actionable intelligence from social media data. The first is the massive amount of information shared leading to information overload, and the second is the proliferation of mistaken and inadvertent misinformation [4]. Taking the 2013 Boston bombing as an example, the number of tweets posted reached 44,000/min just moments after the attack [5]. This was simply too much data to consume, even for official services. To exacerbate the problem, many post-mortem reports indicated that much of this information was inaccurate, with on average, only 20% presenting accurate pieces of factual information [6].

In order to combat the emerging phenomenon of information overload and to support better understanding of situations using large amounts of data, there is a growing need to provide systems and tools that can analyse data and provide enhanced insight. One of the approaches that has been suggested is that of creating 'world views' to allow better understanding of a situation. A world view is a cluster of consistent messages that gives a possible view of a scenario [7]. It contains key aspects in support of an individual's perception of environmental elements with respect to time or space; the comprehension of their meaning; and the projection of their status after some variable has changed, such as a predetermined event [8]. By presenting users with a more complete, consistent, and corroborative picture of a situation, the notion of world views can help to enhance a user's awareness in a scenario or situation. We believe that this enhanced awareness can serve as a crucial starting point (i.e., not the full solution) to address the problem of misinformation in social media.

In this paper, we aim to address some of these issues through the development and evaluation of a system supporting user and organisational situation awareness using social media data. The goal of our system is to facilitate the analysis of datasets of multiple posts, and allow the clustering of consistent posts into different world views. These views can provide valuable insight into on-going scenarios (e.g., crises), that could then lead to better decision-making (e.g., where to send emergency responders as in the case of the London Riots [9]). A main research challenge that we seek to tackle here is the creation of a novel system that can *understand* data through the application of Natural Language Processing (NLP) techniques, and determine consistent and corroborated information items. This work is intended to complement our existing efforts of building a suite of tools for individuals and organisations that can allow actionable intelligence to be gained from open source information [10–13]. These tools would be tailored for analysis of online content while also possessing interfaces suited for human cognition.

The remainder of this paper is structured as follows. "Related work" section reflects on the relevant literature in the fields of social media, misinformation, and situation awareness. Next, in "System approach" section we present our world view extraction approach to understanding social media data. Here we also detail the use, application, and scope of the system. We provide an overview of the system architecture in "System implementation" section. In "Evaluation and discussion" section, we report on an evaluation of our system involving a comparison to a number of existing systems that have similar aims. Finally, we conclude the article in "Conclusion and future work" section and consider future work.

Related work

The proliferation of social media has made it a practicable medium to acquire insights about events and their development in environments [14]. The role of social media in natural disaster crisis management became clear during the 2010 Haiti earthquake [15], and has increased significantly since then [16]. The research community has focused on two general areas to gain the most value from social media especially with such situations in mind. These include, tackling misinformation and its spread, and broad approaches to understanding situations.

The misinformation problem

Misinformation can easily spread in a network of people, highlighting the importance of designing systems that allow users to detect false information. In traditional communication media, machine learning and Natural Language Processing (NLP) are often used to automate the process. However, social networking services like Twitter suffer from intrinsically noisy data that embodies language use that is different from conventional documents [16]. On top of this, Twitter restricts the length of the content published, limiting the usefulness of traditional trust factors (such as length of content) as an indicator of information quality [17]. This results in the accuracy of automated methods with social media data being highly limited, and so manual intervention is often required.

One approach taken by Procter et al. [18] to understand widespread information on Twitter used manual content analysis. The approach utilised code frames for retweet content in order to categorise information flows (e.g., a report of an event), using the groupings to explore how people were using Twitter in the corresponding context. By using a variety of tweet codes, categories such as media reports, rumours (misinformation), and reactions were identified. In particular, rumours were identified as tweets where users had published content, without providing a reference (e.g., an external link).

Other approaches have also attempted to use assessments of individual information items (e.g., tweets) using trust metrics [12, 19, 20]. While these provide a rigorous and automated approach, they often require reliable and a good quantity of metadata to make appropriate judgements. Additional attempts at addressing misinformation issues have also sought to train machine learning classifiers. These would assist with the identification of rumours and low-quality information [17]. Again however, these often require the manual annotation of misinformation to help with classification tasks.

Approaches to understanding situations

Some of the earliest work which attempted to understand situations from social media data was by Sakaki et al. [21] who manually defined a set of keywords relevant for the types of events they wanted to detect (earthquake, shaking, and typhoon). They used a Support Vector Machine (SVM) to classify each tweet based on whether it referred to a relevant event (i.e., an event described by any of the keywords) or not. This approach was limited as the set of keywords needed to be defined manually for each event, and hence a separate classifier needed to be trained.

By acknowledging the importance of syntax and semantics, the approach taken by Vosoughi [22] used a Twitter speech-act classifier which exploited the syntactic

and semantic features of tweets. The classifier was successful in analysing assertions, expressions, questions, recommendations, and requests, but it only used a single tweet as the basis for classification. This skewed the cluster towards the initial seed tweet, impacting the overall awareness gained by a situation. However, Vosoughi's research was very successful in predicting the veracity of rumours by analysing (i) the linguistic style of tweets; (ii) the characteristics of individuals spreading information; and (iii) the network propagation dynamics. The veracity of these features extracted from a collection of tweets was then generated using Hidden Markov Models (HMMs) [22].

Yin et al. propose a system that uses a variety of techniques to cluster tweets in an effective way [16]. It works by initially gathering data in the data capture component, processed using burst detection, text classification, online clustering, and geo-tagging, and then visualisation by the system user. A crucial aspect of the methodology used by Yin et al. is the clustering component, which motivates our research. Specifically, no prior knowledge of the number of clusters is assumed. Within the domain of crisis management, this is especially important as crises evolve over time. This means that partition clustering algorithms such as *k*-means, along with hierarchical clustering algorithms (which require a complete similarity matrix) are not appropriate in this system. Instead, the system extends group average clustering (GAC), an agglomerative algorithm which maximises the average similarity between document pairs [23].

In a similar vein, Suffix Tree Clustering (STC) [24] has been used in Web search engines to group information [25]. In particular, the suffix tree algorithm is well-suited to domains where the number of words in each document is very small [26], such as in social media platforms as Twitter.

Janruang et al. [27], expand STC approaches by proposing an algorithm named Semantic Suffix Tree Clustering (SSTC) which utilises a subject verb object (SVO) classification to generate more informed names in clusters for web search results. The main difference and novelty of our approach is the use of Ukkonen's online, linear time, and space-efficient algorithm [24]. SSTC uses a less space efficient algorithm to ensure all phrases fully appear in the suffix tree [27]. Furthermore, unlike SSTC, our approach will add a trustworthiness assessment to the final clustering process. As SVOSSTC was designed to extract world views from social media information, understanding the reliability and dissemination of information is critical. Thus, SSTC and SVOSSTC both use SVO classification and WordNet as a similarity metric, but SVOSSTC uses a different semantic suffix tree construction algorithm and assesses clusters' trustworthiness. The applications of STC to create world views from social media data have yet to be fully explored, since SSTC was used for web searches, and they serve as a point of interest for our research.

The notion of world views has not been analysed at large by academics. Despite this, one successful approach for using world views to better understand situations was proposed by Rahman [7], who used a number of tactical situation objects (TSOs) to create a set of internally consistent clusters (world views), ranked by an initial provenance metric. TSOs represent an intermediate form of a tweet, allowing natural language to be encoded into a structured XML for the system to process. However, the use of TSOs encoded as XML objects in order to structure tweets is a very simplified

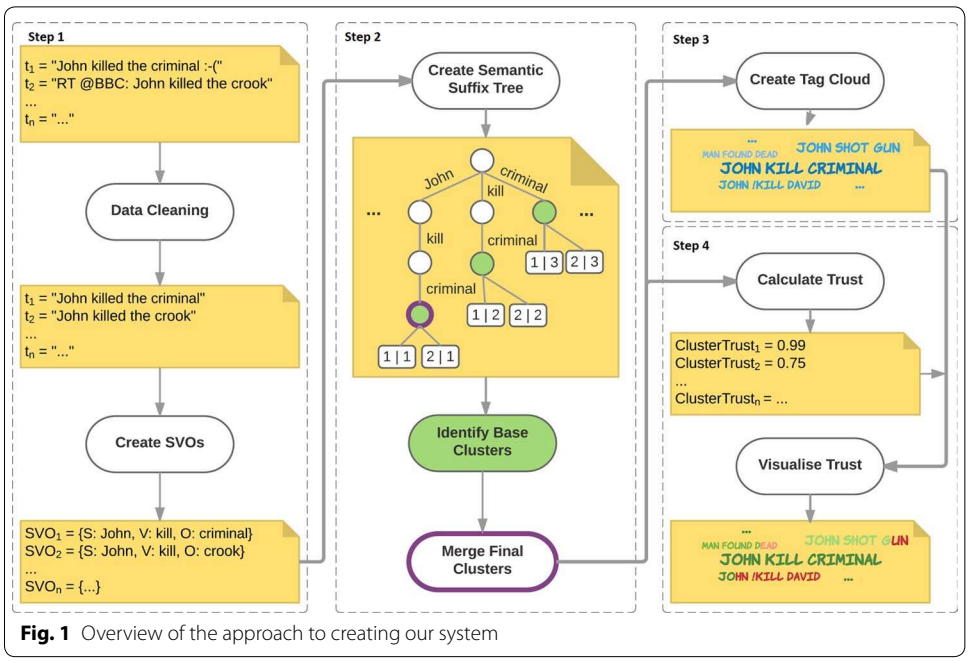

Fig. 1 Overview of the approach to creating our system

assumption in an unstructured real-world environment such as Twitter. Other recent literature has also identified the difficulties of understanding, summarising or assessing this problem [28, 29]

System approach

The goal of our system is to support users in the analysis of multiple posts, and allow the clustering of consistent posts into different views of a scenario (i.e., world view). The approach that we have taken to fulfil this aim consists of four main steps, which are shown in Fig. 1.

In what follows, we present these steps in further detail and explain how they work towards addressing our aim.

Standardising data with Subject–Verb–Object tuples

Data standardisation is challenging with social media, in particular on Twitter, because (i) tweets are noisy and ambiguous; (ii) there is no well-defined schema for the various events reported via Twitter; and (iii) it is not trivial to extract information from unstructured text. We believe that an approach using the Subject–Verb–Object (SVO) typology could be a viable solution, as it is often considered to be the dominant sentence structure in social media communication [30].

The SVO representation [31] is a linguistic structure where the subject comes first, the verb second, and the object third. SVO languages, such as English, almost always place relative clauses after the nouns they modify and adverbial subordinators before the clause modified. An example of the SVO linguistic structure applied to a basic phrase (**P1**) is shown below:

P1: *"David ate lunch"* $\Rightarrow \{(S : David, V : ate, O : lunch)\}$

As discussed in "Related work" section, there have been many attempts to extract value from social media data by using an entire tweet as the input corpus to a system. However, the noise, structure, and complexity of social media data often mean that it is not well suited to the task of information extraction. The SVO representation is useful as it alleviates the issue of data standardisation by producing structured tuples of linguistic information. By using the SVO representation in our approach, we aim to partially address the issue of unstructured data in social media. Below we present a real tweet (**T1**) and show its SVO representation:

T1: *"RT @ABC: The FAA issued a flight restriction"*

$\Rightarrow \{(S : FAA, V : isssued, O : restriction)\}$

In our approach, we apply a series of data cleaning functions to preprocess the tweet before converting it to an SVO representation. These functions include (i) syntax cleaning to reduce inflectional and derivationally related forms of a word to a common base form; (ii) tweet cleaning to remove hashtags, retweets, and other discourse found on Twitter; and (iii) slang lookup to identify unfamiliar words and convert them to standard dictionary English. Each tweet is then parsed into n SVO tuples, along with an identifier corresponding to the original tweet. For example, in **T2** shown below, we have two SVO tuples corresponding to each of the possible SVO representations of the tweet:

T2: *"New images show suspect: Massachusetts police released several images"*

$\Rightarrow \{(S : images, V : show, O : suspect), (S : police, V : released, O : images)\}$

As each tweet in our system is an ordered tuple of length three, as opposed to a long sequence of unordered keywords, our clustering methodology is able to produce more succinct and relevant cluster labels.

We have also considered how lexical databases such as WordNet [32] can enhance our understanding of each of the components in the SVO representation. WordNet contains groups of synonyms, called synsets, which record the relationship between the members of the set. By identifying the synsets from each component of the SVO representation, our approach exploits semantic similarity in order to reduce the future overlap of semantically equal (but syntactically different) cluster labels. Hence, this approach allows us to produce semantically consistent clusters.

The novelty of our approach lies in the analysis of how effective the SVO representation can be in structuring language in order to extract valuable meaning from data. While there is always a potential that data might be lost in any automated information extraction approach, we note that a significant proportion of tweets typically fit into the SVO structure [30]. Consequently, we are therefore motivated by the potential usage of this technique in creating concise and meaningful descriptions for different world views. It is these world views that we aim to utilise to increase situation awareness.

Applying Suffix Tree Clustering to social media data

A suffix tree is a compressed trie (also known as a digital tree) containing all possible suffixes of a given text as the keys, and their position in the text as the values [33]. Suffix trees are particularly useful as their construction for a string S takes time and space linear in the length of S. We define, for a string $S \in \Sigma*$ and $i, j \in \{1, \ldots, |S|\}, i \leq j$, the substring of S from position i to j as $S[i, j]$, and the single character at position i as $S[i]$. The suffix tree for the string S of length n is defined as a rooted directed tree with

edges that are labelled with nonempty strings and exactly m leaves labelled with integers from 1 to j. This is such that: (i) each internal node other than root has at least two children; (ii) no two edges out of one node have edge labels beginning with the same character; and (iii) for any leaf i, the concatenation of the path labels from root to leaf i is $S[i, m]$.

Since such a tree does not exist for all strings S, a terminal symbol that is not seen in the string (usually denoted "$\$id$") is appended to the string. This ensures that no suffix is a prefix of another, and that there will be n leaf nodes, one for each of the n suffixes of S. A slight variation of the suffix tree, used in our approach, is a Generalised Suffix Tree (GST) [33]. A GST is constructed for a set of words instead of single characters, which makes it much more effective for the purposes of understanding sentences, and hence situations. Each node of the GST represents a group of documents and a phrase that is common to all of them. The best algorithm for suffix tree construction is Ukkonen's algorithm [24] which is linear time for constant-size alphabets. In our approach, we adapt Ukkonen's algorithm in order to apply it our GST representation.

Once constructed, the suffix tree approach allows several operations to be performed quickly. For instance, locating a substring in S, locating matches for a regular expression pattern, and many more. However, suffix trees can also be utilised for the purposes of clustering. In particular, Suffix Tree Clustering (STC) has been widely used to enhance Web search results, where short text summaries (also known as 'snippets') are clustered [26, 34]. The STC algorithm groups input documents according to the phrases they share, on the assumption that phrases, rather than keywords, have a far greater descriptive power due to their ability to retain relationships and proximity between words. We aim to exploit this in our approach in order to produce highly descriptive cluster labels.

The clustering methodology has two main phases: base cluster discovery and base cluster merging. In the first phase, each of the documents (tweets) are built into the GST, where each internal node forms an initial base cluster. The second phase then builds a graph representing the relationships between the initial base clusters using a similarity measure. This measure is defined as the similarity in document sets. Effectively, this criterion requires that each cluster must have the most specific label possible to avoid unnecessary, less descriptive, but semantically identical clusters. The clusters are then merged into final clusters if they satisfy the similarity measure. An example of a STC implementation for Web search results is Carrot2 [35].

An advantage of STC is that phrases are used both to discover and to describe the resulting groups, achieving concise and meaningful descriptions. Furthermore, methods that utilise frequency distribution often produce an unorganised set of keywords. We overcome this issue by using STC in our approach. However, some previous attempts at using STC have been limited. For example, if a document does not include any of the exact phrases found in other documents then it will not be included in the resulting cluster, even though it may be semantically identical. Acknowledging this, our approach combines semantic reasoning through utilising WordNet synonym rings (synsets) as part of our SVO tuples (tweets), as mentioned in "Standardising data with Subject–Verb–Object tuples" section, in order to alleviate the issue of semantic consistency. Therefore, our approach introduces *Subject–Verb–Object Semantic Suffix Tree Clustering* (SVOSSTC).

To enable our SVOSSTC approach, we relax the suffix tree constraint that each internal node other than root has at least two children, and enforce the constraint that each label must only be a single word. Using this, we build upon the successful use of STC in Web search results, by exploiting the distinct similarity between snippets and tweets, in order to trial the effectiveness of STC for social media data. Combining the SVO approach outlined in the section above with Semantic Suffix Tree Clustering (SSTC), we propose the following algorithm in our wider approach:

1. Generate SVO representation for each tweet (as previously defined);
2. Create a *Subject–Verb–Object Semantic Suffix Tree* (SVOSST) T with a single root node;
3. For each SVO, ascertain the associated WordNet synset for each part of the SVO representation;
4. For each word in the SVO representation, if the overlap between the synset of the current word and the synset at node n is ≥ 1, then we create a link in T, else we add the synset of the current word to T;
5. After each SVO has been added, we insert a label which includes the tweet identifier (terminal symbol) and the starting branch for the feature word to T;
6. Let each subtree $T_0, T_1, \ldots, T_i$ be a concept cluster, and each node a cluster that has a set of documents to be a member;
7. Each base cluster is then formed using a post-order traversal of the nodes along with the corresponding label;
8. Merge base clusters c_i, c_j with $|c_i \cap c_j| = |c_a|$, then delete c_a, or $|c_i \cap c_j| = |c_b|$, then delete c_b; and
9. Output final clusters.

The novelty and contribution of our work is the definition of an approach which draws on, and extends, existing work from other fields. It applies SVOSSTC to social media data, exploiting the STC algorithm's low complexity and successful applications in Web search results. In particular, our contribution improves upon the basic STC algorithm by using semantic information provided by lexical resources such as WordNet. This increases the likelihood that semantically consistent clusters will be created. Due to the SVO structure utilised in our approach, the scope of semantic similarity is narrowed, ensuring that the cluster labels are as descriptive as possible. For example, traditional approaches to analyse semantic similarity in STC do not consider the structural formation of the sentence, and hence increase both the complexity of the algorithm and obfuscation of the output by incorrectly considering the semantics of structurally different words [36]. The combination of the SVO representation and STC is crucial to our approach, as it produces non-contradictory clusters with a precise semantic meaning. This, as will be discussed below, offers several advances for users in gaining a better situational awareness, especially in situations of crisis.

Creating world views with tag clouds

A tag cloud, or word cloud, is a visualisation technique for textual data, where size, colour, and positioning are used to indicate characteristics (such as frequency and

prominence) of the words. Tag clouds are often used to display summaries of large amounts of text, especially with regard to providing a quick perspective of a situation [13]. In our approach, we use tag clouds as the primary visualisation mechanism to display world views as output from our SVOSSTC approach. We follow key visual design principles for situation awareness [37] to increase their usefulness.

Specifically, we aim to (i) use a filtering mechanism to alter the level of granularity when clustering; (ii) utilise standardised vocabularies with the SVO representation; (iii) highlight the correlation between elements with STC; and (iv) provide a flexible pathway for exploring related information. The approach taken allows users to see both the SVO representation for highlighting general trends, as well as an in-depth overview of all the available tweets that create each cluster. One key difference is that the words in our tag cloud are not independent from each other, but are considered as tags of ordered tuples. Therefore, the data is contextualised into a wider picture, which is an important feature for enhancing situation awareness.

Analysing trust in world views

In order to address the problem of misinformation in social media data, as discussed in "Related work" section, we would need to consider how the world views can be used in the context of trustworthiness. A potential solution to the problem of misinformation was proposed in our earlier work [38], by introducing the notion of information-trustworthiness measures. There, we acknowledge world views as a valuable mechanism in a wider system that can analyse the trustworthiness of information.

Trustworthiness can be considered as an extension of quality (and hence, value) in social media data [39]. It is also perceived that trustworthiness is the likelihood that a piece of information will preserve a user's trust in it [38]. Therefore, our work also uses quality and trust metrics to assess the social media content, and then, based on the values attained and world views produced, informs users of the trustworthiness of the content.

Figure 2 shows how we used coloured tag clouds of the 2011 UK Riots Twitter dataset to convey trustworthiness. Here, letters within words were coloured according to the trustworthiness of the contexts in which they appear, with green being the

Fig. 2 Tag cloud generated from the 2011 London riots [13]

most trustworthy and red the least. From our experimentation with this visualisation approach, we were able to show that it is useful in decision-making, particularly in facilitating a quick, helpful, and accurate overview of a situation [13].

We incorporate the notion of trust into our current work on world views by first creating a world view, then assessing how trustworthy each world view is based on trustworthiness assessment approaches. There are a variety of different trust factors that we considered in our approach, for instance, competence of source, location, user relationships, corroboration, timeliness, and popularity [40]. In the context of our work, we focus on corroboration as a key trust metric from the perspective of world views.

Corroboration is a measure of how many different sources agree with the content of the information provided. Traditionally, clustering engines are limited in respect to corroboration for a number of reasons. Most notably, they can ignore negation in words, resulting in semantically inconsistent clusters. Other approaches cluster a range of tweets and take the size of the cluster as the sole representation of corroboration. This approach should be treated carefully because users may be retweeting misinformation, which falsely identifies high corroboration amongst sources.

Therefore, our approach analyses corroboration, and generally the trustworthiness of a world view by considering three main factors. These are: (i) how much information is corroborated i.e., how large the clusters are (excluding retweets); (ii) identifying the extent to which there are trusted (predefined) parties that can increase the cluster's trustworthiness if present; and (iii) looking to further corroborate clusters with external entities (e.g., news reports on websites) to automatically see which clusters may be more trustworthy. As a start, instead of using real-time identification of related entities, we have decided to use a predefined list of news agencies present on Twitter that we assume (for this case) are trustworthy, and then use these in (iii) to corroborate the trustworthiness within each cluster. We acknowledge that news sources are not always correct or timely, therefore only use this as an extension to our system that may add value. Examples of agencies we have currently included are ABC, BBC News, CBS News, CNN and Reuters. We postulate that a larger proportion of news agencies within a cluster results in a large corroboration, and hence potentially, a higher level of trust. Users of our system would be free to include their own list of pre-defined trusted parties. We use the following algorithm in our approach:

1. Generate world views using the algorithm that we have proposed earlier;
2. Let *clusters* = set of all world views, and $n_i = |cluster_i|$ (without retweets);
3. Let $t = |$trusted news agencies$|$;
4. For each $cluster_i \in clusters$, where $i = 1 \ldots y$ and $y = |clusters|$, sort the clusters from largest to smallest. Then, $x_i = |cluster_i|/\max(n_i)$, where $cluster_i \in clusters$;
5. For each $tweet_k$ in $cluster_i$, observe if it is from a news agency as defined, and let $t_i = |$ trusted news agencies in$cluster_i|$. Then, $c_i = t_i/t$;
6. After each x_i and c_i has been established, calculate $s_i = (0.5 \times x_i) + (0.5 \times c_i)$, to give an output between 0 and 1; and
7. Output $s_i, \forall i \in clusters$.

As can be seen above, we assign equal weighting (0.5 each) to the relative size of the cluster (x_i) and the relative number of news agencies in the clusters (c_i). While this value is somewhat arbitrary, our motivation for these weights is due to both being viewed as equally important factors in measuring corroboration. This is the sole instance of weights usage within our method. In future work, we could seek to examine what weights may be the most appropriate for use, or even opening weighting as an option for users to configure.

It is important to note that our approach is only one of many ways that the problem of misinformation can be addressed. For instance, we could apply the trust metrics to each tweet (as done in [41, 42]) in each cluster and combine these to produce a trustworthy rating per world view. Or, we may look to use machine learning approaches to determine the veracity and credibility of the information (as intended in [17, 43]). The contribution and scope of our work derives instead from extracting, creating, and visualising world views. Our approach explores a new domain by focusing on social media content, rather than the existing encoding formats that are widely used in crises. To take this further, we would blend a range of open-source and closed-source intelligence in order to create a more complete picture for public users of our system or official situation responders.

System implementation

The system architecture is split in three main components, each implemented in Python [44]. These components are: `DataCleaning`, `Clustering`, and `Visualisation`. Each component is an identifiable part of the larger program, and provides a discrete group of related functions. By developing the system in a modular way, we were able to define clear interfaces which are crucial for the extensibility of the system in the future.

Data cleaning

The `DataCleaning` component uploads, processes, and cleans social media data using a variety of pragmatic techniques that we have developed, combined with packages from the Natural Language Toolkit [45] (i.e., libraries and programs for symbolic and statistical NLP for the English language). The three main modules that we created to assist in this task are: the `SyntaxCleaner`, which performs activities such as escaping HTML characters, removing unnecessary punctuation, and decoding the data to the ASCII format; the `TweetCleaner`, that removes URLs, emoticons, and Twitter-specific discourse such as mentions and hashtags characters; and the `SlangLookup` which is responsible for converting colloquial abbreviations such as *"how'd"* and *"m8"* to formal English (how did and mate).

The `DataCleaning` component ensures that a tweet is translated to a (clean) natural language representation before the `SVO` function begins processing. The original tweets are also stored along with pointers from the beginning of each word in a clean tweet to its position in the original tweet. This enables the system to display both versions of the text to enhance readability and understanding. In the next phase, the system processes the clean tweet to obtain the SVO representations.

Once the system has obtained an SVO representation, it uses the `Lemmatiser` and `VerbPresent` functions to further standardise the output for future clustering

operations. In the `Lemmatiser` function, each component of the SVO tuple is transformed with a light stemming algorithm which utilises the `WordNetLemmatizer` [45] function defined in the `NLTK` module. This function reduces inflectional endings to their base or dictionary form using the word itself and the corresponding part-of-speech tag (from `spaCy` [46]) to establish the correct context of the word.

The `VerbPresent` function operates on the verb within each SVO, ensuring that it is in the present tense (e.g., *"gave"* → *"give"*). This is to ensure that the system increases the likelihood of establishing a consistent semantic relationship between the original tweet and SVO representation. The `VerbPresent` function uses the `NodeBox` [47] linguistics library, which bundles various grammatical libraries for increased accuracy. Finally, we repeat this process for each tweet in the input file of tweet objects, and then pass the output to the `Clustering` component.

Clustering

The `Clustering` component ensures that each of the SVOs input from the `Data-Cleaning` component are built into a suffix tree. Once constructed, the SVOs represented by the suffix tree are merged to form world views. The `Clustering` component is also responsible for executing our initial trust metric evaluation; calculating a level of trust which the system associates with each world view.

Firstly, the `Clustering` component passes each SVO to the `SuffixTreeConstruction` module in order to seed the construction of a suffix tree. We build the suffix tree by using the semantic similarity between each component of the SVO tuple, in effect, creating what we define as a *Subject–Verb–Object Semantic Suffix Tree* (SVOSST). Simultaneously, we construct the suffix tree through an on-depth and on-breadth process based on Ukkonen's suffix links [24]. Initially, the suffix tree root is created and the first SVO is taken from the stack. The system then iterates through each of the constituent parts of the SVO tuple (subject, verb, and object), traversing the tree in order to find an overlap between the current word's synonym ring (synset) and the synset at each node. This is implemented using the WordNet [45] corpus from the `NLTK` module. The system utilises the `synsets` function to retrieve the synset for each word, before adding it to the suffix tree. The complexity of SVO algorithm is $O(n)$, where n is the number of words. Therefore the algorithm is linear in the number of words that needs to be processed and we do not anticipate any scalability issues when tested in big datasets.

Next, the `MergingPhase` module produces the final clusters which form our world views. In this function, the system uses each of the base clusters that have been identified by the `SuffixTreeConstruction` function, and decides how to cluster each of them using STC. More specifically, the function executes an implementation of the STC algorithm [25], that we introduce as SVOSSTC. Our system asserts the maximum granularity of the clusters presented in the world view. This is because it is possible to produce base clusters with both a generic (length one) label, and a specific (length three) label with identical document sets. For example, the `ClustSim` function facilitates the avoidance of producing the output *"David kill John"* and *"John"*, when all of the documents stored within the *"John"* cluster are the same as *"David kill John"*.

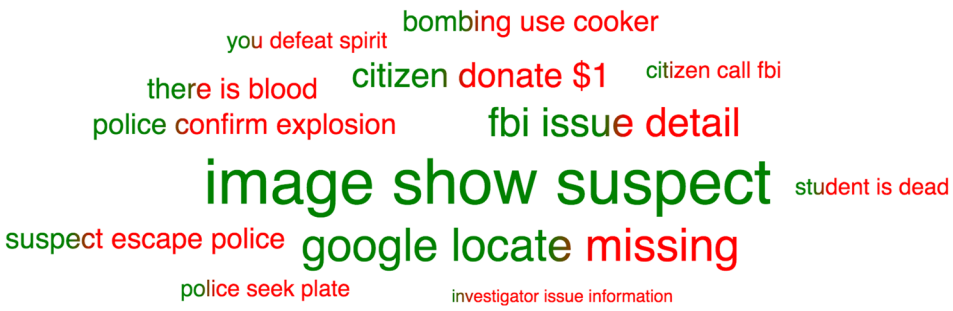

Fig. 3 Tag cloud produced by the `Visualisation` component

Finally, the `ClusterTrustworthiness` component, analyses each tweet within a world view in the context of trusted sources in order to establish a measure for corroboration. A cumulative total of trusted entities is then recorded for each cluster, and divided by the total number of trusted sources that have been defined. The result of this calculation is then combined with the original cluster cardinality, excluding retweets, in order to give an indication of the overall trust of each cluster. This part of the `Clustering` component is highly extensible, providing an interface for various trust metrics to be applied to world views produced by the system.

Visualisation

The `Visualisation` component presents the information generated by the system to the end-user. This component forms the front-end of the system that is responsible for conveying our world views in the form of tag clouds. Figure 3 shows a tag cloud produced by the system based on a subset of real-world data from the 2013 Boston bombing. The tag cloud forms the main contribution of the `Visualisation` component.

The `Visualisation` component uses the score produced by the `ClusterTrustworthiness` function in order to produce a coloured tag cloud. In the tag cloud, the size of the phrase is proportional to the cardinality of the cluster that the phrase represents, and the colour variation (between trustworthy in green, and misinformation in red) highlights possible misinformation identified. Therefore, the cluster *"image show*

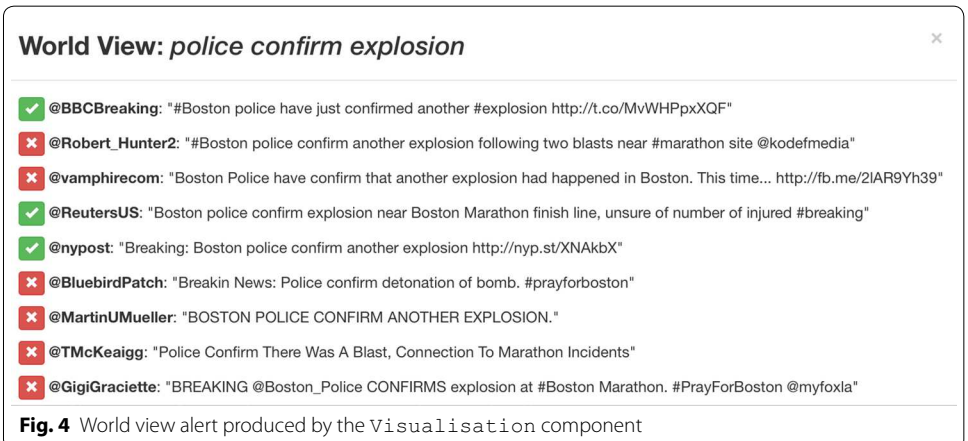

Fig. 4 World view alert produced by the `Visualisation` component

suspect" in Fig. 3 shows a large cluster of information that contains only trustworthy sources corroborating each other.

An in-depth overview of each world view is also generated in the `Visualisation` component when the corresponding element is selected from the tag cloud. The alert window (modal) we have developed is shown in Fig. 4. This modal provides a detailed insight into the composition of the selected world view, and the trustworthiness score associated with each constituent element. Figure 4, presents the world view containing all of the tweets from the *"police confirm explosion"* tag shown in Fig. 3.

Evaluation and discussion

Method

The main research challenge that we seek to tackle in this paper is the creation of a novel system that can understand datasets and determine consistent and corroborated information times. To evaluate our system in the context of this aim, and thereby assess its initial ability to allow an enhanced situational awareness in real-life scenarios, we used three datasets from a variety of crises that have occurred in recent years. In each of these crisis scenarios, Twitter was deemed to play a critical role in the efforts of emergency responders attempting to understand the situation. The datasets used are:

- Dataset 1 (**D1**): 172 randomly gathered tweets from the 2016 Paddington train station incident, caused by an individual standing on a bridge above the railway line [48];
- Dataset 2 (**D2**): 255 randomly gathered tweets from the 2013 Boston bombing, where two pressure cooker bombs exploded killing innocent civilians [49]; and
- Dataset 3 (**D3**): 584 randomly gathered tweets from the 2016 Ivory Coast beach resort attack, where a number people were killed by a gun attack [50].

At this stage in our research, our aim was to assess the effectiveness of our proposed system and also to incorporate the judgement of a human agent in the assessment. Considering this point, we decided to use smaller dataset samples instead of emphasising large datasets, which would transfer the focus to scalability. Moreover, we do not seek to evaluate the trust component of the proposed system, noting our defined research challenge. These are areas that will form central components of our future work in the space.

In summary therefore, five existing systems that use a variety of approaches to cluster information will be applied to **D1**–**D3**, in order to determine the top three world views produced (based on the cardinality of the cluster).

As mentioned in "Related work" section, methods which required manual classification of tweets or manually crafted lists with relevant words for events, such as the SVM approach proposed by Sakaki et al. [21], were not considered relevant for comparison to our work. Our decision was informed by the fact that the system we propose in this paper does not assume prior knowledge of events. In a similar vein, classifiers which aim to identify assertions, questions and expressions based on an initial tweet, such as the one proposed by Vosoughi [22] are relevant for establishing tweets' stance but do not cluster tweets. Therefore, we decided to benchmark our system with approaches that utilise STC, *k*-means and Latent Dirichlet allocation (LDA)

algorithms and whose code was publicly accessible. Our main criteria for selecting *k*-means and LDA clustering with cosine similarity, Carrot2 and Carrot2 STC are the fact that these algorithms do not require prior knowledge and that all three attempt to cluster tweets into different world views.

The systems used are:

- System 1 (**S1**): *k*-means clustering with cosine similarity;
- System 2 (**S2**): Carrot2 with Lingo [51];
- System 3 (**S3**): Carrot2 with Suffix Tree Clustering (STC);
- System 4 (**S4**): Human agent using manual clustering techniques; and
- System 5 (**S5**): LDA clustering with cosine similarity.

The S1, S2, S3 and S5 clusters were selected as they represent the most important information likely to affect an individual's situation awareness. Furthermore, by reducing the scope of evaluation to three world views it is possible to analyse more influential (and critical) information in greater detail.

The first approach we evaluate our system against is the *k*-means clustering algorithm (**S1**). **S1** is widely used in many document clustering approaches [52, 53]. The algorithm performs iterative relocation to partition a dataset into clusters, locally minimising the distance between the cluster centres (centroids) and the data points (tweets represented as vectors). **S1** calculates the distance between each vector (tweet) using the cosine similarity measure [54]), subsequently clustering the vectors (tweets) based on proximity to the centroid.

Carrot2 is an open-source clustering engine that automatically organises collections of documents into thematic categories. Lingo (**S2**) is the first of the Carrot2 algorithms that we evaluate against our approach. **S2** is notable as it reverses the traditional clustering pipeline by first identifying cluster labels and then assigning documents to the labels to form the final clusters. It achieves this using latent semantic indexing (LSI), which analyses the relationship between documents by assuming that similar words will occur in similar pieces of text [55]. **S2** exploits concepts found in the document through LSI, rather than identifying literal terms that are syntactically identical.

STC (**S3**) traverses a suffix tree in order to identify words that occur in more than one document, under the assumption that common topics are expressed using identical sequences of terms. Each of these words gives rise to a base cluster where the nodes contain information about the documents in which each phrase appears. The base clusters are then merged using a predefined threshold, retaining only those documents that meet a predefined minimal base cluster score.

The fourth approach used is a human agent (**S4**). In order to establish a fair comparison, we decided to recruit an individual that had experience with social media and could understand clustering techniques. Specifically, the person selected to cluster the tweets into different world views was a final year Trainee Solicitor studying for a Graduate Diploma in Law. Our decision was justified because we believed that this individual would use typical analytical reasoning to extract the world views.

The final approach we utlise is LDA, which assumes that documents represent mixtures of topics [56]. Each topic is modeled as a probability distribution over a set of words. Thus, given a fixed number of topics (denoted by n), LDA creates a "probabilistic graphical model" [56] that estimates the probability of a topic's relevance to a particular document (i.e., $p(topic|document)$) and the probability of a word appearing given a particular topic (i.e., $p(word|topic)$). This allows LDA to compute the probability that a particular topic is relevant to a given document. We use the LDA implementation in the Python library gensim;[1] cosine similarity measure [54] is used to determine to ensure reproducibility and consistency, we used the same random seed when performing LDA on each dataset.

Our evaluation considers an objective human agent as an important benchmark in order to understand how the problem of clustering may be approached without the use of complex clustering systems. In order to conduct this evaluation, **S4** manually identifies a number of concepts present in the datasets. In the scenario where more than three concepts are identified, **S4** then reprocesses the candidate clusters in order to propose the top three clusterings (based on the cardinality of each set). We envisage that this approach will give us high accuracy for identifying the ground truth in the evaluation datasets.

Before **D1**–**D3** can be evaluated by **S1**–**S5**, there are a variety of data formats that must be produced from the original data. For *k*-means clustering for instance, we had to implement an approach to translate **D1**–**D3** to vectors of weighted word frequencies. A similar process was performed for LDA algorithm. Carrot2 on the other hand requires input documents to be in the Carrot2 XML format [57].

The evaluation will use **S1**–**S5** to analyse **D1**–**D3** to identify how well each system performs, including the one presented in this paper, and the quality of the world views created. We also present a quantitative comparison of *k*-means, LDA and SVOSSTC where the focus is on the number of cluster created by these algorithms rather than the content and we discuss our results.

Results for qualitative analysis

Tables 1, 2, 3, 4 and 5 show the results of **S1**–**S5** applied to **D1**–**D3**. A variety of results have been obtained, including keyword clusters (i.e., world views) that provide a high level overview of a situation, granular low-level clusters, and clusters that identify negation and semantic similarities.

In what follows, we discuss the qualitative results of the evaluation in order to analyse the success of *Subject–Verb–Object Semantic Suffix Tree Clustering* (**SVOSSTC**) at identifying world views.

Discussion on qualitative analysis

Previous research, along with our findings, suggest that there is still a significant need to utilise a variety of techniques to improve situation awareness. We believe that our approach (**SVOSSTC**) makes a significant step forward in achieving this goal by

[1] https://pypi.org/project/gensim/.

Table 1 Results using *k*-means clustering with cosine similarity (S1)

	S1: ***k*-means clustering with cosine similarity**
D1: Paddington	**C1.1.1**: Run, Unable, Follow
	C1.1.2: Station, Tube, @DailyMirror
	C1.1.3: Closed, Breaking, Jump, Threatening
D2: Boston	**C1.2.1**: Continued, Crossed, Finish, Line
	C1.2.2: Arrested, @BostonGlobe, Terror
	C1.2.3: Eludes, Shuts, Hunt
D3: Ivory Coast	**C1.3.1**: Terrorist, @News_Executive, Seaside
	C1.3.2: Guns, Machine, Gunmen, @DailyMirror
	C1.3.3: Witnesses, Way, @AFP

Table 2 Results using Carrot2 with Lingo (S2)

	S2: ***Carrot2 with Lingo***
D1: Paddington	**C2.1.1**: Paddington due to emergency services dealing
	C2.1.2: Services are currently unable to run
	C2.1.3: Service between Edgware Road and Hammersmith due
D2: Boston	**C2.2.1**: News
	C2.2.2: Released
	C2.2.3: Blood
D3: Ivory Coast	**C2.3.1**: Shooting
	C2.3.2: Ivory Coast beach resort
	C2.3.3: Hotel in an Ivory Coast resort popular

Table 3 Results using Carrot2 with STC (S3)

	S3: ***Carrot2 with STC***
D1: Paddington	**C3.1.1**: Dealing with incident, Emergency services dealing, @GWRHelp
	C3.1.2: Incident at Royal Oak, Police incident, Due to a police
	C3.1.3: Station
D2: Boston	**C3.2.1**: Explosion, Boston Marathon
	C3.2.2: Bombing
	C3.2.3: Victims, Blood, Run
D3: Ivory Coast	**C3.3.1**: Breaking, Shooting
	C3.3.2: Beach
	C3.3.3: Reports, Beach resort, Resort in Ivory Coast

providing a foundation for information credibility, and through blending syntax and semantics in the context of document clustering.

It was crucial to our evaluation that the systems (**S1**–**S5**) overcome the limitations discussed in "Related work" section, by using the datasets (**D1**–**D3**) despite the information source, composition, and quantity. Through conducting the evaluation, there were four key themes that we identified, which affected the level of understanding an individual can obtain in a situation: (i) the level of granularity that each clustering method

Table 4 Results using a human agent (S4)

	S4: ***Human***
D1: Paddington	**C4.1.1**: No train service
	C4.1.2: Man jumping off a bridge
	C4.1.3: Several delays
D2: Boston	**C4.2.1**: Marathon explosion
	C4.2.2: People give blood
	C4.2.3: Images released showing suspect
D3: Ivory Coast	**C4.3.1**: Terrorist attack on beach hotel
	C4.3.2: Gunmen armed with machine guns
	C4.3.3: Lots killed

Table 5 Results using LDA clustering (S5)

	S5: ***LDA clustering***
D1: Paddington	**C1.1.1**: services, an, with, emergency, dealing, incident, not, running
	C1.1.2: the, a, royal, at, oak, on, are, police
	C1.1.3: to, royal, police, at, and, a, oak, due
D2: Boston	**C1.2.1**: boston, to, the, marathon, of, in, suspect, bombing
	C1.2.2: to, in, bombing, marathon, hospital, suspect, is, boston
	C1.2.3: boston, marathon, suspect, explosion, the, bombing, at, police
D3: Ivory Coast	**C1.3.1**: in, beach, resort
	C1.3.2: in, hotel, ivory
	C1.3.3: in, resort, ivory

provides; (ii) the way in which negation is dealt with; (iii) the way in which syntax and semantics are used; and (iv) the impact of large quantities of data. The remainder of this section will discuss these themes.

Granularity of clusters

Handling granularity is of crucial importance for delivering appropriate information to users, enabling them to match their information needs as accurately as possible [58]. The need for granularity is enhanced when a variety of information sources are being combined into clusters, in order to avoid oversimplification of key points of information. However, some systems struggle with issues of generalisation, which **SVOSSTC** attempts to handle more effectively in the context of crisis management.

In **S1**, it is evident that the system oversimplifies each incident (as shown in Table 1) by specifying a variety of keywords for each cluster label. As **S1** uses individual keywords for its clustering methodology, as opposed to phrases used with **SVOSSTC**, it is difficult to construct descriptive clusters whose labels have semantic dependency upon one another. Instead, little context is provided as to the situation described by **D1**–**D3**. Most notably, when **S1** clusters the Paddington dataset (**D1**) to produce the cluster *"Closed, Breaking, Jump, Threatening"* (**C1.1.3**), the only context we have about the root cause of the situation (where a man is jumping off a bridge) is the singular word *"Jump"*.

In contrast, when **C1.1.3** is compared with the cluster *"Man jump bridge"* (**CF.1.2**) produced by **SVOSSTC**, it is possible to clearly identify the difference in granularity. **SVOSSTC** demonstrates the ability to identify crucial contextual information within **D1**, thus outperforming the basic approach taken in **S1**. In the context of supporting an individual's situation awareness, this world view exemplifies how important granularity is.

The issue of granularity is further exemplified in **S3** which creates the most general labels, most notably *"Dealing with incident"* (**C3.1.1**) and *"Incident at Royal Oak"* (**C3.1.2**). Both examples restate the overarching theme of **D1**, but do not provide any additional contextual information. One reason for this is that **S3**'s thresholds for clustering are crucial in the process of cluster formation. These thresholds are inherently difficult to tune, which results in variable results on some datasets. Furthermore, STC's phrase pruning heuristic tends to remove longer high-quality phrases, leaving only the less informative and shorter ones [59]. We believe that through using the SVO representation it is possible to create world views that contain more structured and contextual information with **SVOSSTC**. This is supported by **CF.1.1**–**CF.1.3**.

Both **S2** and **S3** experience issues with granularity when analysing **D2**. Firstly, **S2** produces clusters which contain single keyword cluster labels *"News"*, *"Released"*, and *"Blood"* (**C2.2.1**–**C2.2.3** respectively). Interestingly, **S2** is unable to capture the relationship between the components of certain tweets. In particular, that it was the *"News"* being *"Released"*. This is also an issue present in **S1**, where relationships amongst words are not considered. In contrast, **SVOSSTC** was able to produce *"Police confirm explosion"* (**CF.2.3**), which is in fact the news that is being released by the police within **D2**.

Furthermore, when **S3** is applied to **D2**, it produces high level keyword descriptors for the resulting clusters. For example, *"Victims, Blood, Run"* (**C3.2.3**) is comparable to the often poor clusters produced by **S1**. This keyword style of cluster labelling is frequent in **S3**, due to the lack of semantic similarity between keywords. In **SVOSSTC**, by using the SVO representation, we believe it is possible to identify semantic similarity in a structured framework. **SVOSSTC** ensures that all possible tweets are added to the target cluster if they have semantic equivalence. This is discussed further in "Syntax and semantics" section.

Similar conclusions can be drawn for LDA clustering. Table 5 reports the most probable words when running LDA with $n = 3$ on each individual dataset of tweets. For some

Table 6 Results using our approach (SVOSSTC)

	SVOSSTC: ***Our approach***
D1: Paddington	**CF.1.1**: Service !run Paddington
	CF.1.2: Man jump bridge
	CF.1.3: Police storm platform
D2: Boston	**CF.2.1**: Image show suspect
	CF.2.2: Google locate missing
	CF.2.3: Police confirm explosion
D3: Ivory Coast	**CF.3.1**: Gunman attack resort
	CF.3.2: Army evacuate beach
	CF.3.3: Gunfire leave dead

topics—such as Paddington and Boston—LDA successfully isolates words that indicate different world views; for example, when run on the Paddington dataset, two topics mention the "*royal police*" and another includes "*services not running*." However, the bag-of-words model inhibits the readability of LDA results, and, when compared to Tables 4 or 6, LDA outputs qualitatively inferior summaries that do not represent a coherent set of non-overlapping world views. For example, a human agent may report "*No train service*", and SVOSSTC yields "*Service !run Paddington*", which is significantly clearer than a bag of words document that includes "*services*" and "*not running*". The findings of the evaluation, with respect to the level of granularity, demonstrate the ability for **SVOSSTC** to produce world views that outperform **S1**, **S2**, and **S3**.

Handling negation

The role of negation has been acknowledged as important in the field of linguistic analysis [60]. In application domains such as sentiment analysis, this phenomenon has been widely studied and is considered to be crucial to the methodology [61]. However, understanding and addressing the role of negation in clustering systems is often overlooked. As negation is a common linguistic structure that affects the polarity of a statement, it is vital to be taken into consideration when clustering information. Furthermore, in the context of situation awareness it is not possible to satisfy Endsley's three levels of situation awareness (perception, comprehension, and projection) [8] without negation. This is because to perceive an environment in the correct context, in order to comprehend the interpretation of these perceptions and subsequently project upon them, negation must be present to enable polarity to be considered.

There are numerous examples of unordered keywords representing cluster labels in the systems evaluated. In short, these systems ignore the basic concept of negation. **S1** using the Paddington dataset (**D1**) is a good example of this phenomenon. It produces the cluster "*Run, Unable, Follow*" (**C1.1.1**) highlighting how unordered keywords, especially those affecting phrase polarity, cannot be mixed together. Specifically, the decision is left to the end-user to gauge whether "*Unable*" refers to "*Run*" or *"Follow"* (it is in fact the former). The issue with ignoring negation is further exemplified in **C1.1.1**, where the keyword *"Run"* is included in the cluster label produced. However, in **D1** the incident at Paddington has prevented services from running, and therefore *"!Run"* would be a far more representative cluster label (where *"!"* indicates negation). LDA clustering suffers similar fate due to the fact that the algorithm vectorises documents and uses bag-of-words functionality to produce clusters. Therefore, in traditional document clustering methods, such as **S1** and **S5**, the fact that the words may be semantically related and temporally related is not taken into account.

Another issue present in **S1**, **S5** and **S2**, is that of extensive stop word removal. Words affecting phrase polarity can often be removed in these approaches, affecting the overall quality of world views produced. When observing the performance of **S2** using **D1**, the cluster *"Service between Edgware Road and Hammersmith due"* (**C2.1.3**) suggests that there is indeed a service running. This is a false statement and serves as an example of how stop word removal affects polarity.

In contrast, a human agent (**S4**) deals with negation with success, as it is simple for a human to acknowledge latent phrase structure. Our approach (**SVOSSTC**) is also able

to handle negation with similar levels of success to that of **S4**. In this sense, **SVOSSTC** produces the cluster *"Service !run Paddington"* (**CF.1.1**) which fully encompasses this phenomenon. **SVOSSTC** achieves this result by assigning negation to words identified during the SVO construction phase. It is necessary to retain all candidate stop words in the input, as they are used to generate the correct part-of-speech tags for this SVO processing procedure. This is crucial in the context of crisis management, and could cause significant problems if not addressed by systems producing world views.

Interestingly, **S3** performs slightly better than other existing approaches when analysing **D1**–**D3**. However, this may be due to the variety of words in the dataset itself, as **S3** is unable to deal with semantically related negation. For example, if there were information such as *"There is not a service running at Paddington"* and *"There is no service running at Paddington"*, **S3** would be unlikely to yield the single cluster result (due to *"not"* and *"no"*) that is possible with **SVOSSTC**.

The evaluation, in the context of handling negation, clearly demonstrates the ability for **SVOSSTC** to perform at a similar level to **S4**, whilst outperforming **S1**–**S3** in the context of crisis management.

Syntax and semantics

Traditional clustering algorithms do not consider the semantic relationship between words, which means they cannot accurately group documents (tweets) based on their meaning [62]. Thus clustering is based on syntax alone, which is not sufficient to cluster a large quantity of structurally inconsistent data from a variety of sources. To overcome these issues, it is critical that semantic reasoning is used when clustering tweets, improving the resultant world views when compared to classical methods.

S2 and **S3** highlight some of the semantic issues when analysing the Ivory Coast dataset (**D3**). For example, **S2** produces the clusters *"Ivory Coast beach resort"* (**C2.3.2**) and *"Hotel in an Ivory Coast resort popular"* (**C2.3.3**) which contain exactly the same contextual data, and are highly similar with respect to semantics. Therefore, these concepts should be clustered into a single world view. However, **S2** is unable to produce the desired result as it attempts to identify certain dominating topics, called abstract concepts, present in the search. **S2** then picks only such frequent phrases that best match these topics. This means that recurring and semantically similar phrases may not be clustered.

Our approach (**SVOSSTC**) avoids such issues as highlighted in the two paragraphs above because it does not produce non-SVO cluster labels, and once SVO tuples have been generated it is possible to address semantic similarity in the constituent parts of the SVO representation. For example, **SVOSSTC** produces the cluster *"Gunman attack resort"* (**CF.3.1**), which we perceive to outperform a cluster label simply stating the location of the attack. Furthermore, the issue of semantics can also be seen in **S3** using **D3**. In this system, each of the top three clusters (**C3.3.1**–**C3.3.3**) effectively reduce to the single cluster **CF.3.1** produced by **SVOSSTC**. By extracting the semantic meaning of a phrase using SVO, it is possible to understand that *"Shooting"* (**C3.3.1**), *"Beach"* (**C3.3.2**), and *"Resort..."* (**C3.3.3**) are more easily represented in **CF.3.1**.

A major limitation of **S3**, and STC more generally, is that if a document does not include the phrase which represents a candidate cluster, it will not be included within

that cluster, despite the fact that it may still be relevant [63]. This is corrected using **SVOSSTC**, which takes this into consideration through the use of word nets ("Applying Suffix Tree Clustering to social media data" section).

Another interesting observation arises with **S3** when using the Paddington dataset (**D1**). In this dataset, **S3** cannot identify the similarities between *"Dealing with Incident, Emergency Services Dealing"* (**C3.1.1**) and *"Incident at Royal Oak, Police Incident, Due to a Police"* (**C3.1.2**), despite them relating to exactly the same concept. This exemplifies that syntactic structure is often too strong and does not allow flexibility in the linguistic style of posts in our evaluation datasets. Instead, the clustering methodology produces many overlapping and semantically similar clusters.

However, there still exist several challenges, such as polysemy, high dimensionality, and extraction of core semantics from texts, which **SVOSSTC** needs to address more effectively to fully exploit the value of social media data in the context of syntax and semantics.

Information overload

Large datasets can allow for a much deeper insight into a scenario by providing a comprehensive, in-depth overview. Despite this being useful in understanding situations and making decisions, when there is too much information to process, there can be issues with information overload [4], making the task of document (tweet) clustering increasingly difficult. In order to support an individual's situation awareness, systems must allow end-users to fully comprehend large datasets, to ensure that key themes can be extracted (this relates to our world views point).

The damaging effects of information overload are present in a human agent (**S4**) using the Ivory Coast dataset (**D3**), the largest of our evaluation datasets (584 tweets). **S4** is unable to identify critical concepts that our approach (**SVOSSTC**) was able to extract from **D3**, including *"Army evacuate beach"* (**CF.3.2**). This is an important milestone for **SVOSSTC**, as **S4** is often considered to have a good understanding and perception of different world views in a situation. However, according to the principle of bounded rationality, humans will only explore a limited range of alternatives and will consider a subset of the decomposition principles in order to make the task cognitively manageable [64]. This highlights the need for approaches, such as **SVOSSTC**, that are able to identify world views, to alleviate the informational capacity suffered by **S4**. It is this capacity that often reduces **S4**'s effectiveness in crisis situations when compared to other systems.

In contrast, **S4** performs increasingly better than **S1**–**S3** and **S5** when analysing **D1**–**D3**, as it produces more effective cluster labels. Interestingly, the output of **S4** is similar to **SVOSSTC** for all datasets, demonstrating the success of the system in performing at a near human level. This is opposed to **S1**–**S3** and **S5** which often struggle to produce labels that come with large amounts of information at a granular level.

It is important to note that despite the success in evaluating **SVOSSTC** against **S4**, the findings represent the process considered by a single objective human agent. We do accept the argument that it may be better to use a panel of individuals/judges and use a consensus for this approach. Multiple individuals could remove any bias or unforeseen issues and therefore this should be pursued as an avenue of future work. Generally

however, in the context of information overload in our evaluation, it does demonstrate some of the major successes of **SVOSSTC** which facilitate a close to human clustering methodology.

Results for quantitative analysis

For our quantitative analysis, we focus on how well cluster algorithms categorise tweets and compare these results to the clusters SVOSSTC generates. We therefore focus our analysis on *k*-means and LDA algorithms. Figures 5, 6, 7, 8 and 9 show the results of **S1**,

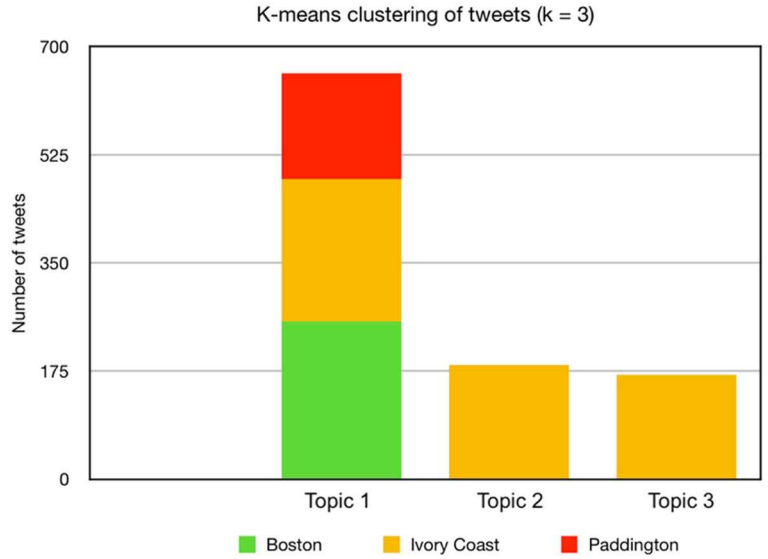

Fig. 5 Output from running k-means on combined dataset of tweets for three clusters, where topic 1 is Boston tweets, topic 2 is Ivory Coast tweets and topic 3 is Paddington tweets

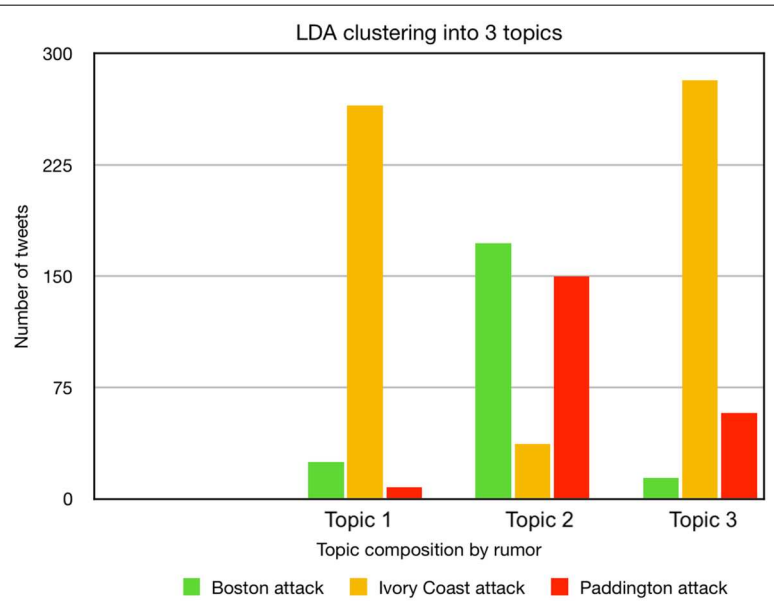

Fig. 6 Output from running LDA on combined dataset of tweets for three clusters, where topic 1 is Boston tweets, topic 2 is Ivory Coast tweets and topic 3 is Paddington tweets

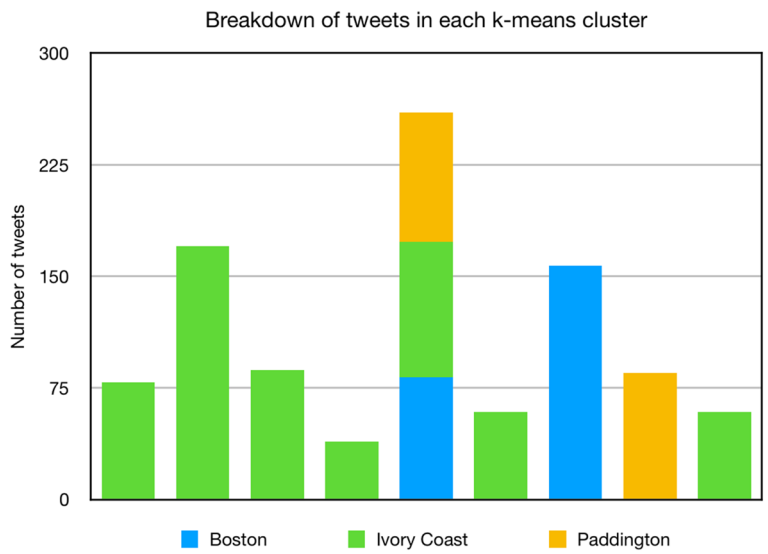

Fig. 7 Output from running *k*-means on combined dataset of tweets for ten clusters

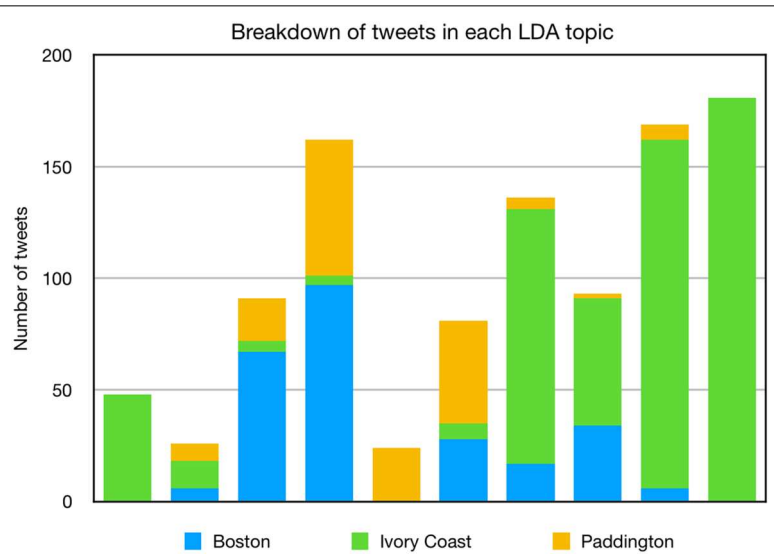

Fig. 8 Output from running LDA on combined dataset of tweets for ten clusters

S5 and **SVOSSTC** applied to a single dataset consisting all **D1**–**D3** tweets. A significant difference between the three algorithms is that **S1** and **S5** require a pre-determined number of clusters where our approach **SVOSSTC** does not. We run LDA and k-means with 3 clusters (the number of different datasets) and with 10 to allow for greater granularity, since datasets **D1**–**D3** comprise different world views (as demonstrated in our qualitative analysis) for the same event.

In what follows, we discuss the quantitative results of the evaluation in order to analyse the success of *Subject–Verb–Object Semantic Suffix Tree Clustering* (**SVOSSTC**) at identifying clusters.

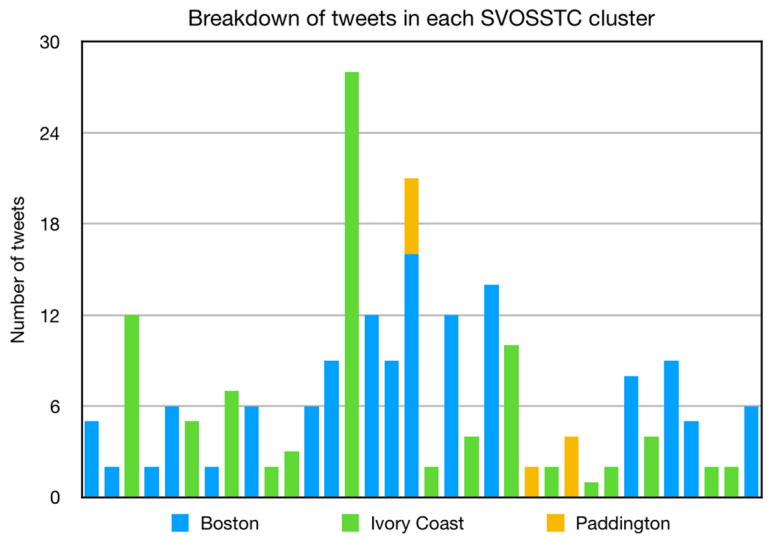

Fig. 9 Output from running SVOSSTC on combined dataset of tweets where number of clusters is not specified

Discussion on quantitative analysis

Our quantitative evaluation emphasises on the accuracy with which clustering algorithms assign tweets to clusters. We create an amalgamation of all the datasets by merging **D1**–**3** into a single dataset. Our aim is to identify whether the clusters created by algorithms are homogeneous ('pure'), thus all tweets derive from the same event (i.e., Boston attacks, Ivory Coast attacks or Paddington incident), or contain tweets from irrelevant events (heterogeneous). The presence of homogeneous clusters indicates that the algorithm is performing well.

Figure 5 clearly demonstrates that k-means algorithm does not separately cluster tweets about the three different incidents. Although Clusters 2 and 3 consist only of tweets relating to the Ivory Coast, Cluster 1 contains all tweets on the Boston Marathon, all tweets on the Paddington incidents, and some tweets from the Ivory Coast attack. One potential explanation for these results is the sheer size of the Ivory Coast dataset; while the first dataset contains 584 tweets regarding the Ivory Coast attack, the Boston dataset contains only 255 tweets and the Paddington only 172.

The inclusion of retweets in the dataset suggests one potential reason for poor clustering behavior. In Fig. 5, 179 of the 185 tweets in Topic 2 are retweets; similarly, 146 of 169 tweets in Topic 3 are also retweets. As k-means does not perform any syntactic analysis, retweets—which are extremely similar under cosine distance—dominate Topics 2 and 3 and likely inhibit qualitative clustering. When increasing the number of clusters to 10, *k*-means performance improves, however a significant number of tweets from all three rumours is still miss-allocated, as Fig. 7 illustrates.

Unlike *k*-means, LDA does not definitively assign a particular document to a topic; instead, it produces a probability distribution representing the relevance of a document to all three topics. In Figs. 6 and 8, we associate a given tweet with the topic LDA assigns the highest probability. Figure 6 illustrates that LDA does not generate

Table 7 Purity measures for LDA, *k*-means and SVOSSTC

	Average purity
LDA	0.77
k-means	0.94
SVOSSTC	0.99

any homogeneous (i.e., high purity) cluster and all clusters contain tweets from all three incidents. Interestingly, when increasing the number of clusters to 10, there is no significant improvement in the performance, as Fig. 8 demonstrates.

When SVOSSTC is run on the combined dataset of tweets, 41 final clusters are generated. While some clusters associate tweets with their retweets, thus resembling *k*-means and LDA behaviour, others effectively congregate world views about a single incident, such as the presence of gunman in a resort in the Ivory Coast. However, SVOSSTC also yields one cluster that does not facilitate greater situational awareness. In this cluster many tweets discussing the Boston Marathon are clustered with others describing the Ivory Coast attack, probably because of the WordNet similarity of words used to describe the two terrorist attacks. We should note that SVOSSTC performs better than *k*-means and LDA, since in these algorithms there are clusters which contain tweets from all three incidents.

An assessment of Figs. 7, 8 and 9 using the purity evaluation measure substantiates our qualitative assertion that SVOSSTC yields finer and more insightful clusters. LDA yields clusters with an average purity of 0.77 (see Table 7), thus reflecting its inability to consistently separate tweets of a particular incident. Seven LDA clusters contain tweets from all three incidents. *K*-means reveals better performance with average purity 0.94. However, although most *k*-means clusters relate to a single incident, one includes dissimilar tweets from all three incidents, including train outages at Paddington, descriptions of the Boston bombing, and reports of gunmen in the Ivory Coast.

SVOSSTC yields average purity 0.99; as described above, although one cluster contains tweets from two separate incidents, it groups tweets that pertain to similar terrorist attacks. Comparison of LDA, *k*-means, and SVOSSTC using the Rand index metric would likely also highlight the benefits of the semantic clustering approach. However, as LDA and *k*-means require choosing the number of topics, and SVOSSTC does not, purity was seen as a more natural performance metric for assessing SVOSSTC against *k*-means and LDA as it does not require computation of a confusion matrix with a fixed number of topics.

Summary

This evaluation focuses on the qualitative and quantitative aspects of how our approach (**SVOSSTC**) performed in comparison to existing systems such as *k*-means clustering, Carrot2 with Lingo, Carrot2 with STC, a human agent and LDA. The findings were generally seen to support **SVOSSTC** as a useful, viable, and effective approach to addressing the issues of generating world views from social media data.

In the discussion above, we analysed the overall success of our approach by identifying four key themes that came from the existing work. From this, a number of notable

research contributions were recognised. Firstly, by using the SVO representation, we were able to establish a level of granularity that was beyond existing systems in the context of crisis management. Secondly, by handling negation as a core concept in our approach, this allowed us to maintain the polarity of tweets, despite the appearance of frequent phrases. Next, by utilising semantics within **SVOSSTC**, our approach facilitated the creation of meaningful clusters without a high level of unnecessary overlap. Finally, we were able to motivate the use of our system by highlighting the ability for it to overcome information overload in larger datasets. All of these problems often affect existing systems.

Focusing on the quantitative analysis, when applied to a dataset that amalgamates tweets from **D1**–**D3**, both k-means and LDA fail to reliably separate tweets into qualitatively useful categories that relate or distinguish the attacks, irrespective of the number of clusters. These results highlight advantages of the SVOSSTC approach.

However, limitations were identified. In particular, the limitations are two-fold: the ability of the system to extract an SVO representation from datasets regardless of the scenario; and the level to which semantic reasoning is an effective tool for overcoming polysemy and extracting core semantics from text. These points also raise questions regarding the refinements of the approach necessary, refinements which may be possible as new and more advanced clustering techniques become available and usable. Furthermore, we intend to explore the use of Word2Vec [65] in the future enhancements of our approach as this may improve the functionality of topic extraction by addressing issues such as polysemy. Word2Vec is a computationally-efficient set of models used to produce word embeddings, which have become popular in the Natural Language Processing field. While we feel that the limitations discovered are important issues, we do not believe that they seriously undermine the contribution of this research and the system proposed.

Overall, from the results of the evaluation it is evident that a majority of the findings were in support of **SVOSSTC** as an approach to facilitate the creation of world views to aid situation awareness.

Conclusion and future work

In this paper, we sought to support user and organisational situation awareness through the research and development of a system to analyse social media data. The specific goal of our system was to facilitate the analysis of datasets of multiple posts, and allow the clustering of consistent posts into different world views. Having identified gaps in existing research, relating to the lack of semantic consideration and a syntactic approach that was too narrowly focused, we settled on a *Subject–Verb–Object Semantic Suffix Tree Clustering* (SVOSSTC) approach in order to produce world views. The advantage of SVOSSTC was found in its ability to create semantically consistent clusters of information with succinct cluster labels, applying techniques observed in the Suffix Tree Clustering (STC) algorithm. Our evaluation supported this advantage as we discovered that a majority of the findings were in support of SVOSSTC, regardless of some caveats to its application.

There are various interesting options for future research. One area is to explore alternative typologies such as Verb–Subject–Object (VSO) and Verb–Object–Subject (VOS)

which are all present in Verb–Object (VO) languages such as English. Furthermore, the inclusion of other VO typologies would enable better identification of data that can be encoded in a structured format (subject, verb, and object). This data can then be semantically clustered with minor alterations to the existing Semantic Suffix Tree Clustering (SSTC) implementation. This would allow for a more complete and consistent picture of a situation due to more representative data being included in the output of the system.

Another avenue of further consideration is the semantic relationship between words in the SVOSSTC algorithm. A slight limitation of synonym rings (synsets) is the fact that the algorithm which links synonymous words together may not contain an exhaustive list of words. Therefore in rare occasions, some words may not be recognised and can be omitted. Instead of focusing entirely on synsets as a measure of semantic similarity, other approaches can be utilised. For example, the use of lexical chains [66] could help to increase the accuracy of semantic clustering. This view is motivated by the characteristics of lexical chains being able to provide a context for the resolution of an ambiguous term, and being able to identify the concept that the term represents. For example, "*Rome \rightarrow capital \rightarrow city*" represents a lexical chain. This demonstrates how concepts in world views could henceforth be semantically clustered in combination with the existing SVOSSTC approach, increasing the probability of finding a semantic overlap between words.

In our further work, we are also keen on experimenting with the variety of new clustering approaches being published (e.g., [67, 68]). These may increase the accuracy of our approach or the efficiency of its use, and thereby add to its suitability as a system for better understanding real-world situations. While we did not concentrate on the efficiency of the proposed approach (in general or as compared to the other algorithms), this will be a key factor in our continued work given that ideally, our system will be used in a 'live' context such as an unfolding crisis scenario. Using this, we expect to focus on large-scale datasets and conduct a range of head-to-head comparisons between the various techniques proposed (and our own improved technique). This would allow us to better investigate its efficiency and performance at addressing the key issues of clustering and world-view analysis.

From a user focused perspective, another area which we could explore in future work is how individuals use and respond to different systems (S1–S4) and the clusters they produce. In the ideal case, we would also look to try this in a real-world event to gain as authentic a response from users as possible. This would provide further validation of our proposed system and its utility. Most importantly, as we look towards building on this research and addressing the issue of misinformation online, we will need to better understand how to dynamically identify trustworthy sources in real-time. Whilst corroborating clusters (world views) with predefined trusted sources is possible, it is not infallible. Trustworthy sources will need to be relevant and updated to consider the context of the scenario and the user of the system. In this way, our approach will be able to make significant progress on addressing the misinformation problem online.

Authors' contributions

This article is the product of research led by CK, in collaboration with JRCN, IA and ABM. As a result, the core research and experimentation was conducted by CK, with JRCN and IA providing direction and guidance during the research; ABM also contributed to the experimentation and comparison of the proposed approach. All parties assisted in the journal manuscript drafting stage. All authors read and approved the final manuscript.

Author details
[1] Department of Computer Science, University of Oxford, Oxford, UK. [2] School of Computing, University of Kent, Canterbury, UK. [3] Stanford University, Stanford, USA.

Acknowledgements
There are no acknowledgements at this time.

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

Availability of data and materials
The Twitter data used in this study is not shareable due to the company's terms of service.

Funding
There are no funding bodies to acknowledge for this research.

Publisher's Note

Received: 11 March 2018 Accepted: 10 July 2018
Published online: 30 July 2018

References

1. Karandikar A (2010) Clustering short status messages: a topic model based approach. Ph.D. thesis, University of Maryland
2. Blandford A, Wong BW (2004) Situation awareness in emergency medical dispatch. Int J Hum Comput Stud 61(4):421–452
3. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World Wide Web. ACM, pp 591–600.
4. Rodriguez MG, Gummadi K, Schoelkopf B (2014) Quantifying information overload in social media and its impact on social contagions. In: Proceedings of the 8th international conference on weblogs and social media. AAAI
5. Withnall A (2013) Twitter uncovers the top tweets of 2013. http://www.independent.co.uk/life-style/gadgets-and-tech/twitter-uncovers-the-top-tweets-of-2013-9007027.html. Accessed 8 Jul 2017
6. Smithsonian (2013) In the wake of the Boston Marathon Bombing, Twitter was full of lies. http://uk.reuters.com/article/uk-ivorycoast-attack-idUKKCN0WF0LB. Accessed 25 Jan 2018
7. Rahman SS (2014) Data fusion for human intelligence and crisis management: handling information from untrusted sources. Ph.D. thesis, University of Warwick
8. Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. J Hum Fact Ergon Soc 37(1):32–64
9. The Guardian (2012) Riot rumours on social media left police on back foot. http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter. Accessed 25 Jan 2018
10. Nurse JRC, Creese S, Goldsmith M, Lamberts K (2012) Using information trustworthiness advice in decision making. In: Proceedings of the international workshop on socio-technical aspects in security and trust (STAST) at the 25th IEEE Computer Security Foundations Symposium (CSF). IEEE, pp 35–42
11. Nurse JRC, Agrafiotis I, Creese S, Goldsmith M, Lamberts K (2013) Building confidence in information-trustworthiness metrics for decision support. In: Proceedings of the 12th IEEE international conference on trust, security and privacy in computing and communications (TrustCom). IEEE, pp 535–543
12. Nurse JRC, Agrafiotis I, Goldsmith M, Creese S, Lamberts K (2014) Two sides of the coin: measuring and communicating the trustworthiness of online information. J Trust Manag 1(1):5
13. Nurse JRC, Agrafiotis I, Goldsmith M, Creese S, Lamberts K (2015) Tag clouds with a twist: Using tag clouds coloured by information's trustworthiness to support situational awareness. J Trust Manag 2(1):1–22
14. Scott J (2012) Social network analysis, 3rd edn. Sage, London
15. Keim ME, Noji E (2010) Emergent use of social media: a new age of opportunity for disaster resilience. Am J Disaster Med 6(1):47–54
16. Yin J, Lampert A, Cameron M, Robinson B, Power R (2012) Using social media to enhance emergency situation awareness. IEEE Intell Syst 27(6):52–59
17. Castillo C, Mendoza M, Poblete B (2011) Information credibility on Twitter. In: Proceedings of the 20th international conference on World Wide Web. ACM, pp 675–684
18. Procter R, Vis F, Voss A (2013) Reading the riots on Twitter: methodological innovation for the analysis of big data. Int J Soc Res Methodol 16(3):197–214
19. Cho J-H, Chan K, Adali S (2015) A survey on trust modeling. ACM Comput Surv 48(2):28
20. Nurse JRC, Agrafiotis I, Creese S, Goldsmith M, Lamberts K (2013) Communicating trustworthiness using radar graphs: a detailed look. In: Eleventh annual international conference on privacy, security and trust (PST). IEEE, pp 333–339
21. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web. ACM, pp 851–860
22. Vosoughi S (2015) Automatic detection and verification of rumors on Twitter. Ph.D. thesis, Massachusetts Institute of Technology
23. Yang Y, Pierce T, Carbonell J (1998) A study of retrospective and on-line event detection. In: Proceedings of the 21st annual international conference on research and development in information retrieval. ACM, pp 28–36

24. Ukkonen E (1995) On-line construction of suffix trees. Algorithmica 14(3):249–260
25. Ilic M, Spalevic P, Veinovic M (2014) Suffix tree clustering. In: Proceedings of the 23rd international conference on electrotechnical and computer science. pp 15–18
26. Branson S, Greenberg A (2002) Clustering Web search results using suffix tree methods. Stanford University, Final Project Report 1:32
27. Janruang J, Guha S (2011) Semantic suffix tree clustering. In: First IRAST international conference on data engineering and internet technology, DEIT. Citeseer
28. Rudrapal D, Das A, Bhattacharya B (2018) A survey on automatic Twitter event summarization. J Inf Process Syst 14(1):79–100
29. Kwon A-R, Lee K-S (2013) Opinion bias detection based on social opinions for twitter. J Inf Process Syst 9(4):538–547
30. Schrading N, Alm CO, Ptucha R, Homan C (2015) #WhyIStayed, #WhyILeft: Microblogging to make sense of domestic abuse. In: Proceedings of the 14th annual conference of the North American Chapter of the ACL: Human Language Technologies. ACL, pp 1281–1286
31. Tomlin RS (1988) Basic word order: functional principles. J Linguist 24(1):213–217
32. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11):39–41
33. Löbhard C (2008) Suffix trees. http://wwwmayr.in.tum.de/konferenzen/Jass08/courses/1/loebhard/Loebhard_Paper.pdf (**unpublished report**)
34. Zamir O, Etzioni O (1998) Web document clustering: a feasibility demonstration. In: Proceedings of the 21st annual international conference on research and development in information retrieval. ACM, pp 46–54
35. Stefanowski J, Weiss D (2003) Carrot2 and language properties in Web search results clustering. In: Proceedings of the 1st international Atlantic web intelligence conference on advances in web intelligence. Springer, pp 240–249
36. Dang Q, Zhang J, Lu Y, Zhang K (2013) WordNet-based suffix tree clustering algorithm. In: Proceedings of the international conference on information science and computer applications
37. Lanfranchi V, Mazumdar S, Ciravegna F (2014) Visual design recommendations for situation awareness in social media. In: Proceedings of the 11th international conference on information systems for crisis response and management. pp 792–801
38. Nurse JRC, Agrafiotis I, Goldsmith M, Creese S, Lamberts K, Price D, Jones G (2015) Information trustworthiness as a solution to the misinformation problems in social media. In: Proceedings of the 1st international conference on cyber security for sustainable society. pp 28–35
39. Kelton K, Fleischmann KR, Wallace WA (2008) Trust in digital information. J Am Soc Inf Sci Technol 59(3):363–374
40. Nurse JRC, Rahman SS, Creese S, Goldsmith M, Lamberts K (2011) Information quality and trustworthiness: a topical state-of-the-art review. In: Proceedings of the international conference on computer applications and network security (ICCANS)
41. Nurse JRC, Creese S, Goldsmith M, Rahman SS (2013) Supporting human decision-making online using information-trustworthiness metrics. In: Proceedings of the international conference on human aspects of information security, privacy, and trust. Springer, pp 316–325
42. Gupta A, Kumaraguru P, Castillo C, Meier P (2014) Tweetcred: real-time credibility assessment of content on twitter. In: International conference on social informatics. Springer, pp 228–243
43. Giasemidis G, Singleton C, Agrafiotis I, Nurse JRC, Pilgrim A, Willis C, Greetham DV (2016) Determining the veracity of rumours on Twitter. In: Proceedings of the 8th international conference on social informatics. Springer, pp 185–205
44. Python Software Foundation: Python 2.7.12 documentation. http://docs.python.org/2.7/. Accessed 25 Jan 2018
45. NLTK Project: NLTK 3.0 documentation. http://www.nltk.org/. Accessed 25 Jan 2018
46. spaCy GmbH: spaCy: Industrial-strength natural language processing. http://spacy.io/docs/. Accessed 25 Jan 2018
47. Experimental Media Research Group: NodeBox: Create visual output with Python programming code. http://www.nodebox.net/code/index.php/Linguistics. Accessed 25 Jan 2018
48. Gutteridge N (2016) Police storm London train station following reports of man on overhead bridge. http://www.prnewswire.com/news-releases/web-users-increasingly-rely-on-social-media-to-seek-help-in-a-disaster-100258889.html. Accessed 25 Jan 2018
49. Eligon J, Cooper M (2013) Blasts at Boston Marathon kill 3 and injure 100. http://www.nytimes.com/2013/04/16/us/explosions-reported-at-site-of-boston-marathon.html. Accessed 25 Jan 2018
50. Penney J, Aboa A (2016) Al Qaeda gunmen kill 16 in Ivory Coast beach attack. http://uk.reuters.com/article/uk-ivorycoast-attack-idUKKCN0WF0LB. Accessed 25 Jan 2018
51. Osiński S, Weiss D (2004) Conceptual clustering using Lingo algorithm: Evaluation on open directory project data. In intelligent information processing and web mining. Springer, Berlin, pp 369–377
52. Ravindran RM, Thanamani AS (2015) *k*-means document clustering using vector space model. Bonfring Int J Data Min 5(2):10
53. Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. In: Proceedings of the KDD workshop on text mining. pp 525–526
54. Ozdikis O, Senkul P, Oguztuzun H (2012) Semantic expansion of hashtags for enhanced event detection in Twitter. In: Proceedings of the 1st international workshop on online social systems. CiteSeerX
55. Dumais ST (2004) Latent semantic analysis. Ann Rev Inf Sci Technol 38(1):188–230
56. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
57. Osiński S, Weiss D. Carrot2 user and developer manual for version 3.14.0. http://download.carrot2.org/head/manual. Accessed 25 Jan 2018
58. Pfennigschmidt S, Voisard A (2009) Handling temporal granularity in situation-based services. International Computer Science Institute, California
59. Weiss D (2001) A clustering interface for web search results in Polish and English. Ph.D. thesis, CiteseerX
60. Wiegand M, Balahur A, Roth B, Klakow D, Montoyo A (2010) A survey on the role of negation in sentiment analysis. In: Proceedings of the workshop on negation and speculation in Natural Language Processing. ACL, pp 60–68
61. Lapponi E, Read J, Øvrelid L (2012) Representing and resolving negation for sentiment analysis. In: Proceedings of the 12th international conference on data mining workshops. IEEE, pp 687–692

62. Wei T, Lu Y, Chang H, Zhou Q, Bao X (2015) A semantic approach for text clustering using WordNet and lexical chains. Exp Syst Appl 42(4):2264–2275
63. Zhang D, Dong Y (2004) Semantic, hierarchical, online clustering of Web search results. In: Proceedings of the 6th Asia-Pacific web conference. Springer, pp. 69–78
64. Simon HA (1996) The sciences of the artificial. MIT press, Cambridge, Mass
65. Patterson J, Gibson A (2017) Deep learning:a practitioner's approach. O'Reilly Media, Inc., Sebastopol
66. Barzilay R, Elhadad M (1999) Using lexical chains for text summarization. Advances in automatic text summarization. MIT Press, Cambridge, pp 111–121
67. Liu W, Ye M, Wei J, Hu X (2017) Compressed constrained spectral clustering framework for large-scale data sets. Knowl Based Syst 135:77–88
68. Kozlowski M, Rybinski H (2017) Semantic enriched short text clustering. In: International symposium on methodologies for intelligent systems. Springer, pp 435–445