Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery

WILEY

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

PERSPECTIVE OPEN ACCESS

A Conceptual Framework for Human-Centric and Semantics-Based Explainable Event Detection

Taiwo Kolajo[1,2] | Olawande Daramola[1]

[1]Department of Informatics, University of Pretoria, Pretoria, South Africa | [2]Department of Computer Science, Federal University Lokoja, Lokoja, Kogi State, Nigeria

Correspondence: Olawande Daramola (wande.daramola@up.ac.za)

Received: 18 December 2023 | **Accepted:** 2 September 2024

Edited by: Witold Pedrycz, Editor-in-Chief

Funding: The work is supported by the National Research Foundation (NRF), South Africa, and University of Pretoria, South Africa.

Keywords: event detection | explainable AI | explainable event detection | human-centric explanations | semantic-based explainable AI

ABSTRACT

Explainability in the field of event detection is a new emerging research area. For practitioners and users alike, explainability is essential to ensuring that models are widely adopted and trusted. Several research efforts have focused on the efficacy and efficiency of event detection. However, a human-centric explanation approach to existing event detection solutions is still lacking. This paper presents an overview of a conceptual framework for human-centric semantic-based explainable event detection with the acronym HUSEED. The framework considered the affordances of XAI and semantics technologies for human-comprehensible explanations of events to facilitate 5W1H explanations (Who did what, when, where, why, and how). Providing this kind of explanation will lead to trustworthy, unambiguous, and transparent event detection models with a higher possibility of uptake by users in various domains of application. We illustrated the applicability of the proposed framework by using two use cases involving first story detection and fake news detection.

1 | Introduction

Event detection is a computational operation for automatically identifying significant incidents in a specific place and time. It answers the questions of when, where, what, and by whom (Panagiotou et al. 2016). This notion is no longer adequate since it leaves out important aspects of explainability. For an event to be humanly understandable, it must have *what*, *who*, *where*, *when*, *why*, and *how* (5W1H) dimensions (Miller 2019). A comprehensive definition that fits into the 5W1H notion is that an event is a significant incident happening at a specific time for specific reasons, with associated entities such as humans, objects, and locations (Chakman et al. 2020; Chen and Li 2020). The typical instances of event detection include first story detection, breaking news, anomaly detection, criminal detection, and catastrophe breakout (Cardinale et al. 2022; Hu et al. 2022).

Different approaches, such as rule-based reasoning, machine learning, deep learning, lexicon-based techniques, and hybrid techniques, have been employed in event detection (Kocher and Kumar 2021). However, no crime detection or first story detection method now in use has been able to offer a human-centric explanation (Chi et al. 2017; Kolajo and Daramola 2023). A human-centric explanation is essential for practitioners and users to ensure that machine learning models are widely used and trustworthy. An explanation focused on people must adapt to the user's context, be understandable,

and establish credibility via provenance (Yuan et al. 2023). Building a decision-making process that is more reliable and long-lasting requires event detection systems to incorporate human-centric explainability (Vemula 2022). The trustworthiness, explainability, provenance, and dependability lacking in existing event detection systems will be attained by leveraging semantics for explainable event detection (Kolajo and Daramola 2023). The formulation of the proposed conceptual framework for human-centric semantic-based explainable event detection (HUSEED) is based on a detailed review of the literature to understand the strengths and limitations of existing event detection systems (Kumar et al. 2023) and the affordances of Human-centric AI (Bingley et al. 2023), XAI (Ehsan et al. 2023), and semantic technologies (Donadello & Dragoni, 2021). Event detection systems that evolve from the proposed conceptual framework for HUSEED would be more suited to answer the 5W1H questions and satisfy the desire for human centricity. In this paper, we describe the generic HUSEED framework and demonstrate its applicability by showing how the generic HUSEED framework could be adapted for two common exemplar use cases of event detection: first story detection and crime detection. Thus, the contributions of this paper are:

1. a conceptual framework for human-centric explainable event detection (HUSEED), which, to the best of our knowledge, is the first attempt in this area.
2. an overview of how human-centric and explainable first story detection can be realized.
3. an overview of how human-centric and explainable online crime detection and many facets (such as misinformation, terrorism, computer misuse, fake news, hate speech, and many more—Figure 2) can be realized.

The remainder of the paper is organized as follows. Section 2 is the literature review, comprising theoretical background on relevant aspects: explainable event detection, human-centric explanations, semantics for explainable AI, and a review of related work on first story detection (FSD) and crime detection. The proposed HUSEED framework is presented in Section 3, while Section 4 presents the use cases for the proposed framework. The discussion on the framework is presented in Section 5, while the paper is concluded with a summary in Section 6.

2 | Literature Review

This section presents background on relevant aspects and reviews related work on first story detection (FSD) and crime detection.

2.1 | Explainable Event Detection

Automatic event detection typically starts with machine learning techniques. Deep learning models, typically opaque models called "black boxes," perform better than other models but lack transparency because of their complex algorithms and self-learning. As a result, explainability and performance are compromised (Arrieta et al. 2020). This challenge necessitated the development of XAI. XAI aims to explain the black box machine learning models without sacrificing performance (Machlev et al. 2022).

A fully explainable event detection system must respond to the 5W1H dimension inquiry in a manner that is understandable to humans (Chakman et al. 2020; Chen and Li 2020; Kolajo and Daramola 2023). Without combining domain knowledge with XAI methods, event detection from social media streams cannot be explained in a way that is understandable to humans. Kolajo and Daramola (2023) argue that social media feed characteristics like short messages, grammatical and spelling errors, mixed languages, ambiguity, and poor sentence structure call for the use of semantic web technologies to enhance human comprehension. The existing event detection systems that have tried to provide explanations used the limited information in the social media streams. None of the existing event detection systems has captured the six dimensions of 5W1H to provide explanations (Khan et al. 2023).

2.2 | Human-centric Explanations

Human-centric XAI design is required to provide understandable explanations. In event detection, explainability is crucial for practitioners and users to ensure that results generated by AI models are widely accepted and trusted (Kolajo and Daramola 2023). There will be the need to design and develop more explainable predictive models to achieve the required level of human-centricity in event detection systems. Optimizing predictive models would only be worthwhile if they can solve the human-centric task of providing explanations (Ai et al. 2023).

Event detection systems must incorporate explainability that is human-centric to be able to facilitate a decision-making process that is more reliable and long-lasting (Vemula 2022). To encourage confidence and wider adoption, this enormous gap calls for creating human-centric explainable event detection models (Ali et al. 2023; Garibay et al. 2023).

2.3 | Semantics for Explainable AI

A user-centric semantic-based explanation should be understandable, appealing to the user, adapt to their context, and include provenance (Borrego-Diaz and Paez 2022). XAI solutions should focus on explanations rather than mathematical mappings, and reasoning engines should operate on precise semantics to provide human-centric explanations (Doran, Schulz, and Besold 2017). Logical analysis of ML models is insufficient, as it does not use an explicit relationship between model learnt features and knowledge base concepts. Domain knowledge is necessary for human-comprehensible explanations, and data analysis alone is insufficient (Hind 2019).

Combining semantic web technologies and AI systems is essential for delivering explanations in natural language (Lecue 2020). Linking explanations to ontologies can tailor the degree of specificity and generality to different user profiles

and facilitate effective knowledge transmission (Tiddi and Schlobach 2022).

2.4 | First Story Detection

First story detection (FSD) is a process that identifies previously unheard news and designates it as "old" or "new" as it enters a social media stream (Moran et al. 2016). It is more challenging than topic detection and tracking (TDT), as it requires an online system to find the initial article for every new occurrence (Ghosal et al. 2022). The nearest neighbor algorithm has been suggested as an effective strategy for FSD (Wang 2019). However, they often fail when postings reporting the same event in textually different ways are marked as first stories due to lexical variance. Word embeddings can help identify genuine first stories by extending tweets with semantically connected terms (Egger 2022). Similarly, the 3-nearest neighbor clustering (3-NN) was modified to identify a first story by representing documents as TF-IDF weight vectors (Vuurens and de Vries 2016).

Osborne et al. (2012) proposed a conceptual framework to enhance first story event detection from Twitter using Wikipedia, using the common information between tweets and Wikipedia articles on both a textual and temporal level. Panagiotou et al. (2016) proposed entity-action-entity triplets for FSD. The authors used word2vec to address lexical variations. Time-aware first narrative identification from Twitter streams was proposed by Qiu et al. (2015, 2016). Wurzer and Qin (2018) introduce the k-term hashing FSD model, which compares incoming narratives to prior tales. Wang (2019) also proposed a new term rate (NTR) technique for FSD based on the number of new phrases in a story. Panagiotou et al. (2021) propose an enhanced and scalable FSD framework using named entities and relation extraction.

From the review, none of the existing approaches to FSD has been able to provide a human-centric explanation for FSD. Providing an explanation by answering the 5W1H dimension questions will provide a trustworthy, understandable, transparent, and trustworthy explanation to all categories of users. This makes the need for FSD systems that will provide human-centric explanations compelling.

2.5 | Crime Detection

The rise of user-generated content on social media platforms has led to increased opportunities for crime, which can be committed anonymously and disseminate hostility (Kaur, Singh, and Kaushal 2021). Crimes on social media can be categorized as public disorder (violent disorder), computer misuse (spear phishing, denial of service), terrorism (radicalization), misinformation (Fake news, rumor), fraud (Astroturfing, employment scam, impersonation), violence against the person (stalking and harassment—hate speech, cyberbullying) (Drury et al. 2022). Crime detection techniques can be categorized into rule-based, machine learning, lexicon-based, and hybrid approaches (Rony, Bakchy, and Rahman 2020). Rule-based systems often include manually crafted rules or keyword blacklists, but they can be time-consuming and challenging to create and manage, often generating false positives and negatives.

Fake news is a growing issue, with machine learning being used to detect it, but the models used are often "black boxes." Major events like the 2016 US presidential elections and the COVID-19 pandemic have been fertile ground for disinformation campaigns (Mishima and Yamana 2022). Meel and Vishwakarma (2021) suggested a CNN semi-supervised framework to investigate latent patterns in unlabeled data with fake news detection using the Kaggle dataset. Capsule neural networks were used for false news detection, and different embedding models for different lengths were used (Goldani, Momtazi, and Safabakhsh 2021). An attention-based convolutional bidirectional long short-term memory (AC-BiLSTM) method was developed to detect fake news, with a notable increase in accuracy rate (Trueman et al. 2021).

Feature engineering is a popular method for detecting cyberbullying, expanding the standard text representation of words by adding new features and dimensions based on linguistic cues (Zhao, Zhou, and Mao 2016). Semi-supervised techniques have been proposed to reduce manual annotation work by employing a bootstrapping approach to produce labels for unlabeled data (Xiang, Hong, and Rosé 2012) and confidence in the voting function to extract negative and positive data (Nahar et al. 2014). Unsupervised methods have been developed to effectively cluster documents with bully traces (Capua, Nardo, and Petrosino 2016). Automated identification of cyberbullying on social media has been presented using the word2vec algorithm (Zhao, Zhou, and Mao 2016). Naive Bayes has been found to perform better than other classifiers in detecting abusive language on Indonesian social media (Ibrohim and Budi 2018). Other methods for cyberbullying include fuzzy fingerprints (Rosa et al. 2019) and deep learning (Kumar, Tyagi, and Das 2021).

Deep learning architectures (CNN, FastText, and LSTM) have been used to detect hate speech on the Twitter dataset. Various deep learning-based models like TC-CNN (Gadek and Guelorget 2020), BerConvoNet, Entity recognition, Sentence reconfiguration (Choudhary et al. 2021), Ordinary Differential Equation network (ESODE) (Ma et al. 2021), content-based characteristics and the WOA-Xgbtree algorithm (Sheikhi 2021) have been used to detect hate speech. However, the existing solutions focus on something other than human-centric explainability, which is crucial for user acceptance and adoption. Therefore, a framework for online crime detection systems that offer human-centric explanations is necessary.

3 | Toward Human-Centric and Semantic-Based Explainable Event Detection

AI systems are creating safer, faster autonomous systems, but their effectiveness is limited by their inability to communicate their decisions to humans. To be widely used, event detection models must be reliable, clear, transparent, adaptive, intelligible, and provenance-enabled. Explainable Event Detection (XED) can address these limitations by leveraging semantic technologies like knowledge graphs, domain knowledge, ontologies, and open knowledge repositories (Kolajo et al. 2020; Pesquita 2021).

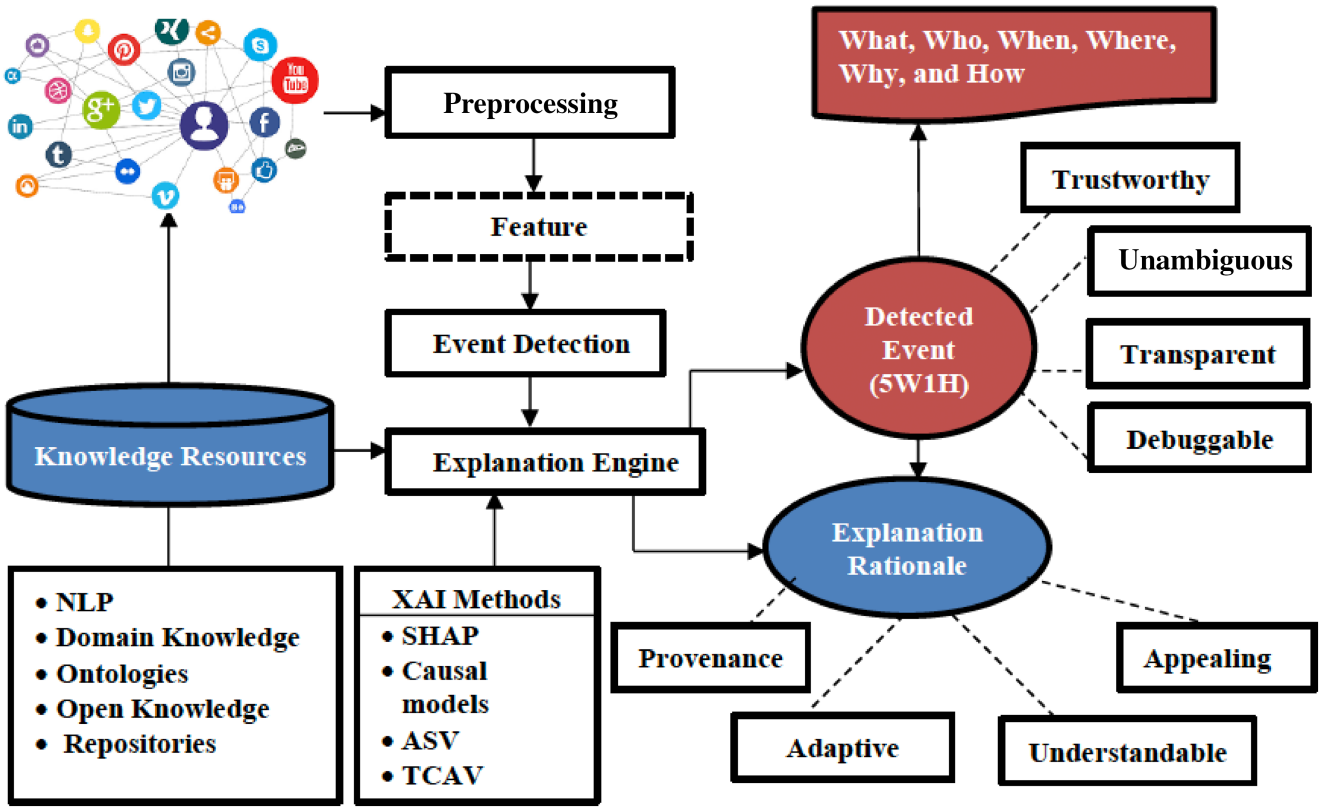

FIGURE 1 | A conceptual framework for HUSEED.

We present how HUSEED can be adapted to achieve human-centric explanation in specific application use cases of event detection, such as first story detection and crime detection.

3.1 | A Conceptual Framework for HUSEED

We argue that the vision of a generic explainable event detection framework (see Figure 1) that possesses the following characteristics is realistic if:

- It is enabled by semantic technologies (ontologies, knowledge graph, open knowledge repositories).
- Offers 5W1H explanations (Who did what, when, where, why, and how).
- That possess human-centric attributes such as adaptive, understandable, appealing, and provenance.

We opine that this will lead to trustworthy, unambiguous, transparent, and debuggable event detection models with a higher possibility of uptake by users in various domains of application.

3.2 | A Component View of HUSEED

The HUSEED framework consists of five main components: data collection, data preprocessing, feature engineering, event detection, and event explanation. Each component of the framework is described next.

3.2.1 | Data Collection

Social media feeds or online news can be collected simultaneously or sequentially alongside data from information sources like Wikipedia or DBpedia, depending on the work at hand. The social media feeds can come from a variety of platforms, including Twitter, Facebook, YouTube, WhatsApp, Instagram, WeChat, TikTok, Sina Weibo, QQ, Telegram, Snapchat, Qzone, Pinterest, Reddit, LinkedIn, Quora, Discord, Twitch, Tumblr, and Mastodon.

3.2.2 | Data Preprocessing

In natural language processing (NLP), it is customary to employ a regular expression to eliminate URLs, Tags, mentions, and non-ASCII characters. This may not be the case for all tasks; for example, for the goal of detecting rumors or fake news, URLs, Tags, mentions, and non-ASCII characters are the linguistic characteristics of the user submissions. The linguistic component is a piece of statistical data that will be mined alongside user postings to determine whether or not they are authentic. Tokenization and normalization are the steps that follow in the data preparation process. The data preprocessing may end here, and the feature engineering step will now get the normalized data. However, the noisy nature of social media feeds may require further preprocessing. Slangs, acronyms, and abbreviations may need definitions depending on the purpose of event detection. To do this, corpora of English terms from the natural language toolkit (NLTK), slang, acronyms, and abbreviations (SAB) identified need to

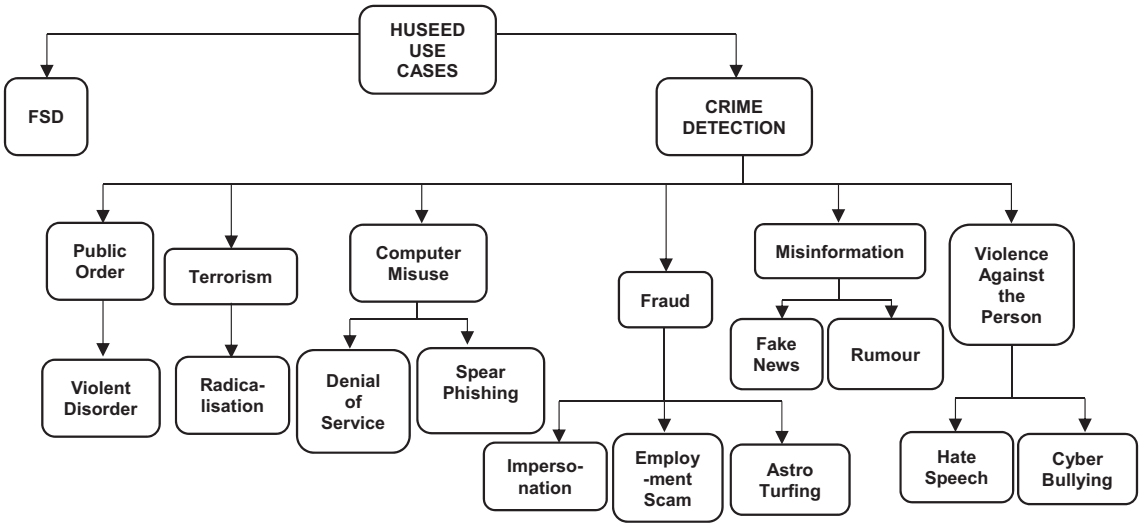

FIGURE 2 | HUSEED use cases.

be filtered from the social media stream. A dictionary of slang, acronyms, and abbreviations may receive the filtered SAB words afterwards. The dictionary is then used to derive the meanings of SAB. Each SAB may have several interpretations; thus, choosing the option with the most appropriate meaning from the available choices is necessary. Kolajo et al. (2020) proposed a typical algorithm for disambiguating SAB terms. After the disambiguation of the SAB terms, the enriched preprocessed social media feed can then be passed to the next stage, feature engineering.

3.2.3 | Feature Engineering

At this stage, the preprocessed social media feed or dataset will be transformed into sets of values essential for event detection. In order to identify the linguistic and temporal similarities and differences between the two streams of information during the data collection stage (for instance, tweets and Wikipedia), the resulting knowledge source and social media feed will be compared. Comparing the knowledge source and social media feed can be used as the provenance of such social media feed. Each social media post is subjected to named entity recognition using an API. Entity extraction is a technique for locating important components in an unstructured text and classifying them into specified groups, such as names of people, places, organizations, quantities, amounts of money, dates, and times. Knowledge sources like DBpedia may be used to explain each entity taken from the social media post. The event cluster summary known as nuggets may also be created using entity extraction. The knowledge source used may be cited in support of the event discovered in social media feeds as provenance. One of the attributes that a detected event should have to be deemed explainable is provenance. This stage can include statistical feature extraction, sentence reconfiguration, and embedding, depending on the specific task.

Features can be selected from the three aspects, including user characteristics involved in propagating the news, news content linguistic features, and propagation network structures. Linguistic features are extracted, which include several positive and negative words, the number of mention symbols, the number of URLs, the number of special characters, and the number of the same words recognized as different entities. User characteristics (gender, age, specialization, dependability, acceptance, authority, number of statuses, account creation date, username, daily post count, and time between posts) will be extracted to obtain additional information about the user. The number of followers, friends, re-posts, and favorites are features of the user propagation network. The original social media post's sentences are rearranged using sentence reconfiguration to produce more regular sentences. A lexical parser can be used to achieve this. Documents will be transformed into a fixed-length vector as part of the embedding task using tools like word2vec, sent2vec, doc2vec, Glove, FastText, and others. The embedded document is passed to the next stage.

3.2.4 | Event Detection

At this stage, the embedded document from the previous stage will be classified into event or non-event. Various methods, depending on the event detection tasks, can be used. Unsupervised, semi-supervised, supervised machine learning, and semantics-based approaches can be used. Of interest here are the documents classified as events. These are the ones that will be passed to the explanation engine to generate an explanation.

3.2.5 | Explanation Engine

At this stage, explanations are generated for the documents classified as events in the previous stage. To provide answers to the 5W1H dimension questions, that is, *who*, *what*, *when*, *where*, *why*, and *how*, entity extraction can be used to provide *who*, *what*, *when*, and *where* explanation by linking these entities to knowledge sources to provide appealing, understandable explanations, and provenance. We are not just looking at the detected event based on the social media post but also verifying such events through knowledge sources. The focus of the *why* and *how* explanation is to understand the rationale and procedure used by the machine learning (ML) algorithm

to reach a decision. For example, knowledge rules can be generated from XAI global explanation methods such as Shapley Additive exPlanations (SHAP), RuleXAI, Causal models, Asymmetric Shapley Values (ASV), Testing with Concept Activation Vectors (TCAV), Layer-wise Relevance Propagation (LRP) (Holzinger et al. 2022). RuleXAI can be used for global and local explanations (Macha et al. 2022). Subsequently, an instance-based learning method like case-based reasoning (CBR) [87] can generate the why and how explanations for decisions made by the machine learning algorithm for specific instances.

4 | Use Cases of HUSEED

This section explores how HUSEED can be adapted to address research gaps in existing event detection models (see Figure 2), focusing on explainable first story detection (XFSD) and explainable crime detection (XCD).

4.1 | Adaptation of HUSEED for First Story Detection

The first story detection (FSD) system must be scalable, accurate, and explainable, answering 5W1H dimension questions. The proposed explainable FSD (XFSD) uses methods from Qiu et al. (2015) and improved versions (2016). The nugget-based method summarizes every event but stores all previous tweets. The XFSD uses parallel knowledge graph sources like Wikidata, DBpedia, and social media posts to compare pages and tweets over textual and time dimensions. The knowledge graph resource serves as supporting evidence for the event detected in social media posts and answers 5W1H dimension questions. The process workflow for explainable first story detection is shown in Figure 3. The specific steps to achieve explainable first story detection (XFSD) are described subsequently.

4.1.1 | Data Collection

Any social media platform of choice can be used for data collection. For illustration purposes, we use Twitter (viz x.com) for the data collection. The FSD uses new tweets as input; only previously received tweets are available when dealing with the current tweet. Using Twitter API, tweets are streamed as input.

4.1.2 | Data Preprocessing

URLs, tags, mentions, and non-ASCII characters are removed from the tweets using regular expressions. Tokenization and normalization are performed using the NLTK library tool. It is necessary to give meanings to slang, acronyms, abbreviations, and irregular sentences. To do this, the slang, abbreviations, and acronyms (SAB) found are sifted from the social media feed involving corpora of English words in the natural language toolkit (NLTK). The SAB terms that have been filtered are then sent to a dictionary of slang, acronyms, and abbreviations to provide the SAB definition. Because of the possible multiple meanings connected to each SAB, there is a need to disambiguate the ambiguous terms and select the best sense from the available meanings.

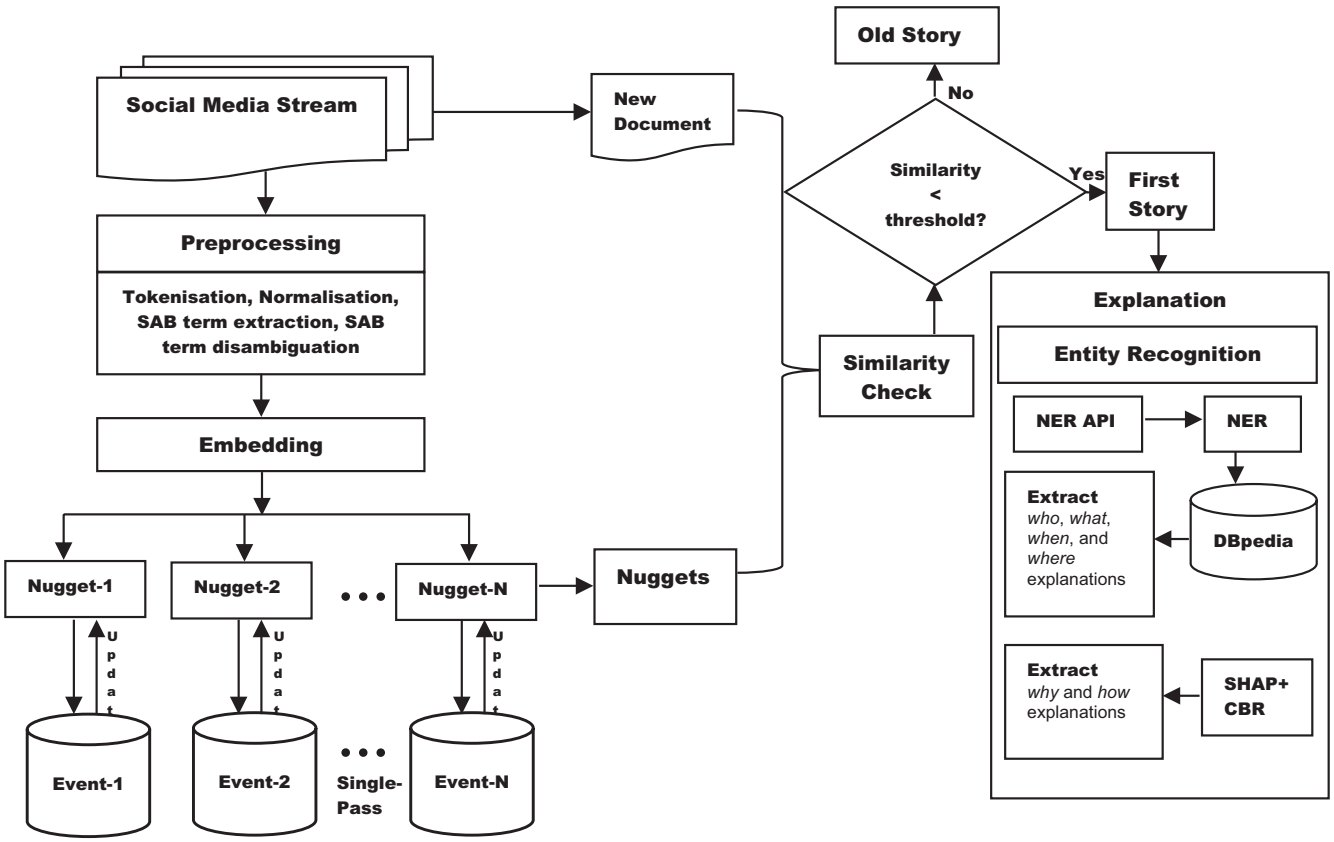

FIGURE 3 | The explainable first story detection.

ALGORITHM 1 | Slang, Acronym, Abbreviation Disambiguation Algorithm (SABDA) (Kolajo et al. 2020).

Input: tweet text, sabt
Output: enriched tweet text
//Procedure to disambiguate ambiguous SAB in tweets by leveraging SABDA over the usage //examples of SAB found in the local vocabulary (*ikb*)
Notations:
sjk: the current tweet being processed
slngs: slangs; acrs: acronyms; abbr: abbreviation
sabt: a collection of slang/acronym/abbreviation terms
w_i: an individual slang/abbreviation/acronym in sabt
st_i: i^{th} usage example of w_i in sabt found in *ikb*

```
procedure disambiguate_slngs/acrs/abbr
    for each wi in sabt
        best_sense=disambiguate_slng/acr/abbr (wi, sjk)
        display best_sense
    end for
end procedure
function extract_usage_examples(wi, ikb)
    if we are in ikb then
        select wi.sti //pick all the usage examples (sti) w.r.t wi
    else
        ikb.not_found = wi
end function

function disambiguate_slng/acr/abbr(wi, sjk)
    usage_senses = extract_usage_examples(wi, ikb)
    //usage_senses: a collection of usage examples of wi in sabt found in the ikb
    //usage_senses → {st1, st2, …stn | m ≧1}

    int[] score //an array of semantic relatedness scores

    for each sti ∈ usage_senses of wi
        for i= 1 to n do
            // n is the total number of usage examples for wi
            score [i] = relatedness(sti,sjk)
        end for
        best_score = max(score[i])
    end for
    STI ← best_score
    return sti
    map sti with defi  //(where defi ∈ definition)
    replace wi in sabt in the tweet with defi
end function
```

The Slang-Acronym-Abbreviation Disambiguation Algorithm (SABDA) proposed by Kolajo et al. (2020) can be used for disambiguation (see Algorithm 1). The enriched preprocessed social media feed is then passed to the next stage, feature engineering.

4.1.3 | Feature Engineering

Term features such as the enriched tweets from the data preprocessing stage will be transformed into a fixed-length vector using doc2vec, a standard document embedding tool with stable text classification performance.

4.1.4 | First Story Detection

Semantic histogram-based incremental clustering (SHC) is a text clustering algorithm that uses embedded tweets to identify event clusters. A nugget is chosen based on local and global significance, with local significance indicating a word's commitment to the tweet's focal idea and global significance relating to its commitment to sub-points (see Algorithm 2). The algorithm saves computational time by comparing new social media posts to the nugget of each event cluster, updating it dynamically for each detected event. When a new document arrives, it will be compared with the nugget of each event. If the similarity meets

ALGORITHM 2 | The Nugget Generation Algorithm (Adapted from Kolajo, Daramola, and Adebiyi 2022).

Input: Tweets in ClusterC, CList
Output: Nugget
Begin
  foreach tweet T in ClusterC
    for each word w in tweet T
      $x \leftarrow count\ w_i\ in\ Tweet\ T$
      $y \leftarrow count\ w_i\ in\ ClusterC$
      // Compute weight of each word in tweet T
      $W(w_i) \leftarrow \alpha_1 \log(1 + x) + \alpha_2 \log(1 + y)$
    end for
    //Compute weight of each tweet T
    $Score(T) \leftarrow \sum_{i=1}^{n} W(w_i)$
  end for
  // Get tweet T with highest score
  foreach Score(T) in Cluster C
    $maxScore \leftarrow ScoreT[0$
    ]
    if ScoreT[i] > maxScore **then**
      maxScore = ScoreT[i]
    end if
  end for
End

the threshold of any of the nuggets, the event is added to the bucket of such a nugget and the nugget is updated. Otherwise, it is taken as the first story of the new event.

4.1.5 | Explanation for First Story Detection

Of interest here are the events that are classified as the first story. First story events will be analyzed using named entity recognition (NER) to identify actors in the first story (who), first story event location (this should not be misinterpreted as user location; we are interested in the location where the first story took place (where)), the first story event itself (what), the first story event time, again this should not be misinterpreted as the time the user posted the event but when the event happened (when). The detected first story events will be passed to DBpedia with queries to extract the answers to 5W1H dimension questions using NER and SHAP. The NER API will extract *who*, *what*, *when*, and *where* explanations from DBpedia, while the *why* and *how* explanations will be provided through case-based reasoning (CBR) by using knowledge rules that form the basis for ML decisions extracted from SHAP. The answers to the 5W1H dimension questions will lead to user-comprehensible explanations of detected events.

4.2 | Adaptation of HUSEED for Crime Detection

The HUSEED framework, adapted for explainable crime detection, has been improved by Ma et al. (2021) by adding entity recognition (NER) and sentence reconfiguration components. These components enhance semantic understanding of texts and adjust word order and frequency. We extended the framework of Ma et al. (2021) to answer the 5W1H dimension questions (see Figure 4) to provide explainable fake news in social media.

4.2.1 | Data Collection

The various APIs for the various social media streams can be used to collect data. Information from knowledge sources will be crawled to help in entity recognition at the feature engineering phase.

4.2.2 | Data Preprocessing

For fake news detection, URLs, tags, mentions, and non-ASCII characters will be used for the linguistic features of the user posts. So, there is no removal of these features in this case. The linguistic feature is part of statistical information that will be mined along with the user posts to detect whether the post is fake. The data preparation operation here would be mainly tokenization and normalization, which can be achieved using the NLTK library tool (Malviya and Dwivedi 2022).

4.2.3 | Feature Engineering

Entity recognition, statistics feature extraction, sentence reconfiguration, and embedding are the four sub-steps of this stage:

Entity recognition: Each social media post can be extracted using the NER API. The clarification sentence of every element that shows up in the post will be crawled from the information

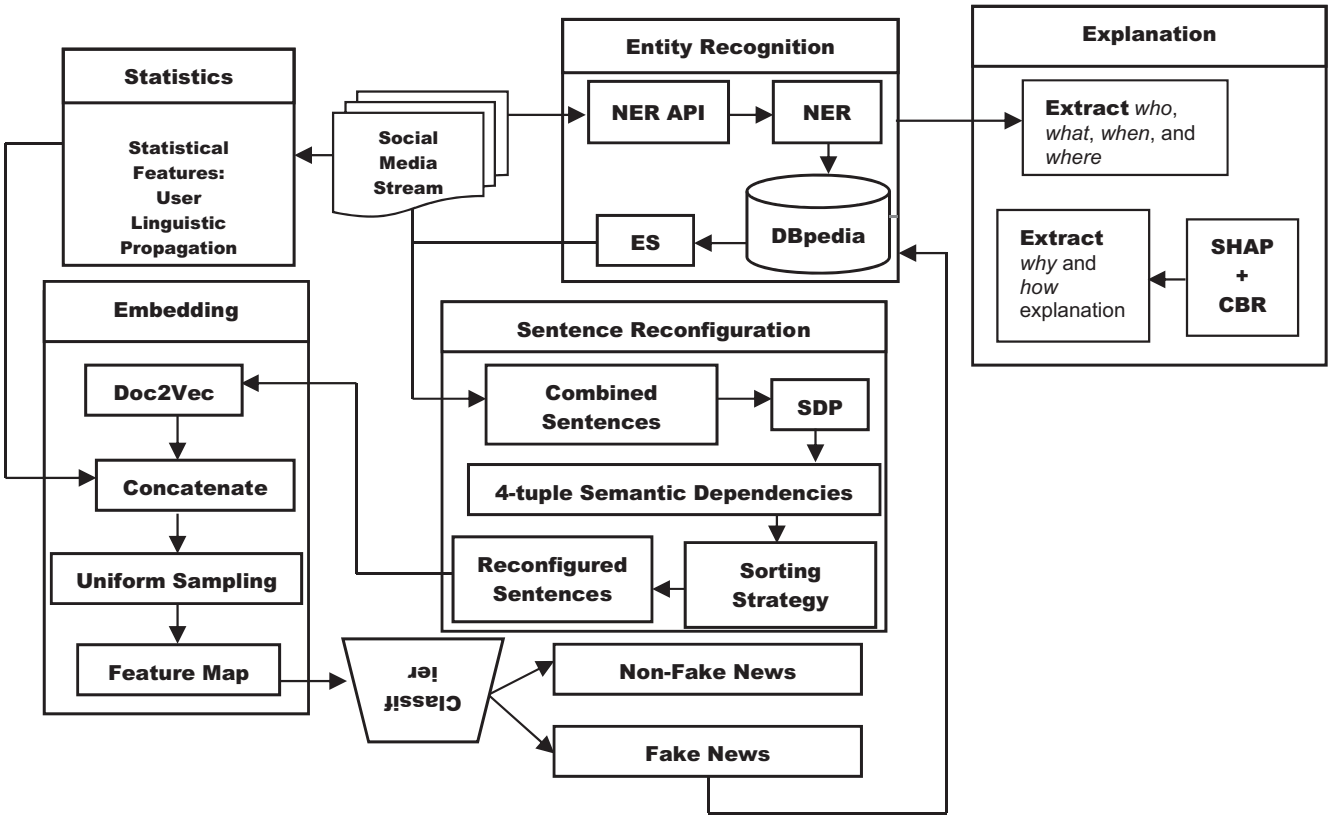

FIGURE 4 | Explainable fake news detection workflow.
Source: Ma et al. (2021).

sources, for example, DBpedia, by utilizing DBpedia Spotlight (Aboucaya, Guehis, and Angarita 2023), which is an entity recognition and name resolution tool, each entity's explanation sentence will be enclosed in square brackets after it.

User characteristics: These are the selected statistical features, which include gender, age, specialization, dependability, acceptance, authority, number of statuses, username, number of daily posts, and time between posts.

Linguistic features: The number of positive and negative words, the number of mentioned symbols, the number of URLs, the number of special characters, and the number of the same words recognized as belonging to various entities are all extracted.

Propagation network structures: These include the number of reposts, the number of favorites, the number of friends, and the number of followers. The Stanford Dependency Parser (SDP) is used to generate a four-tuple sequence for each word in a sentence (stored according to their semantic location (SL) and dependency), resulting in a rearranged sentence. Doc2vec, a document embedding tool, is then used to convert the statistical features and rearrange sentences into a fixed-length vector representing a social media post, comment, or retweet.

4.2.4 | Fake News Detection

The embedded social media posts from the previous step will be passed to text classification algorithms such as CNN, LSTM, and ODE-net. The selected algorithm will be trained to classify the text as fake or non-fake news.

4.2.5 | Explanation for Fake News Detection

Of interest here is the events that are classified as fake news. The fake news events will be analyzed further at this stage to generate an explanation. NER will be used again, but with specific entities such as the propagator of the fake news (who) and fake news event location (this should not be misinterpreted as user location), we are interested in the location where the fake news took place (where), the fake news event itself (what), the fake news event time, again this should not be misinterpreted as the time the user posted the event but when the event happened (when). The detected fake news events will be passed to DBpedia with queries to extract the answers to 5W1H dimension questions using NER and the combination of SHAP and CBR. The NER API will extract *who*, *what*, *when*, and *where* explanations from DBpedia, while the *why* and *how* explanations are provided using SHAP and CBR. Knowledge rules may be generated automatically from the Shapley values using automated knowledge rule extraction algorithms and techniques. For example, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) or FOLD-RM can be used to extract rules from SHAP (Wang, Shakerin, and Gupta 2022). RIPPER is a rule-based learning technique that can be applied to extract knowledge rules automatically (Seerat and Qamar 2015). FOLD-RM provides native explanations for prediction without external libraries or tools. The extracted rules will be used for similarity-based

matching through CBR to generate explanations for specific cases. Providing answers to the 5W1H dimension questions in this way will lead to user-comprehensible explanations and detection of fake news that are consequently trustworthy solutions which can be adopted.

5 | Discussion

To realize HUSEED, the authors assumed that its users could comprehend explanations written or displayed visually. HUSEED aims to improve user comprehension of explanations in natural language. It aims to create reliable, clear, transparent, adaptive, intelligible, and provenance models for event detection. HUSEED will use semantic technologies like knowledge graphs, domain knowledge, ontologies, and open knowledge repositories to overcome the limitations of current event detection methods and improve trustworthiness and transparency. HUSEED can solve the present event detection models' lack of reliability, explainability, trustworthiness, and transparency. HUSEED can get beyond the limits of current event detection methods by utilizing semantic technologies like knowledge graphs, domain knowledge, ontologies, and open knowledge repositories.

HUSEED will combine natural language, visuals, and domain knowledge to provide decision-making insights. However, subjectivity and knowledge acquisition limitations may affect its applicability. HUSEED's explanations may need help to explain context-specific patterns, resulting in limited generalization. Despite these challenges, it will combine human-centricity and semantics to generate context-aware and domain-adaptive explanations.

Automated knowledge rule extraction from XAI methods like SHAP can create interpretable rules for complex machine learning algorithms. Techniques include Decision Tree Induction, Rule-based Learning, Association Rule Mining, Genetic Algorithms, Rule Pruning, Optimization, Feature Discretization, Ensemble Rule Learning, and Visualization (Arsovski et al. 2019; Lundberg and Lee 2017). It is significant to highlight that the effectiveness of automated knowledge rule extraction from SHAP depends on the composition and complexity of the underlying model. Due to the complexity of the model or the dataset, it may occasionally be essential to tradeoff between accuracy and interpretability.

HUSEED is a systematic software development framework aiming to improve initiatives' effectiveness, predictability, and quality. It includes phases, activities, roles, responsibilities, artifacts, deliverables, best practices, and guidelines guiding the creation of software products.

6 | Conclusion

XAI has emerged as a crucial component in event detection systems to meet the requirements of understandability, transparency, interpretability, and trustworthiness. In addition to supporting the established standards, guidelines, and regulations, XAI plays a crucial role in gaining user trust. In this paper, we have discussed the conceptual notion of human-centric and semantic-based explainable event detection. We have argued that the 5W1H dimension questions must be answered for event detection systems to produce explainable, transparent, and trustworthy solutions, which aid their uptake by users and practitioners. Likewise, semantic technology needs to be introduced to the existing event detection frameworks to achieve the goal of human-centricity. Users are more likely to trust and embrace event detection systems with explainability attributes. So far, none of the existing event detection systems currently in use have answered the 5W1H dimension questions. Thus, we proposed HUSEED as a generic, explainable event detection framework that can cater to this deficiency of existing event detection systems. As an exemplar, we also described how HUSEED can be adapted and applied in two use cases: explainable first story detection (XFSD) and explainable fake news detection (a form of crime detection). Thus, we demonstrated how the HUSEED framework can answer the 5W1H dimension questions to achieve human-centric explanations. We show that by using named entity recognition (NER) combined with XAI methods like SHAP and case-based reasoning (CBR), the objective of human-centric explanations can be achieved by utilizing semantic-based methods and knowledge sources like DBpedia. Hence, we presented HUSEED as a general explainable event detection framework that can be used in various event detection scenarios, including terrorism, spear phishing, denial of service, astroturfing, employment scams, impersonation, hate speech, and cyberbullying.

In future work, we shall develop concrete event detection systems for different use cases based on the HUSEED framework and conduct evaluation studies to ascertain their performance and usability from stakeholders' perspectives. In addition, we will explore the possibility of using LLMs to generate explanations for event detection systems.

Author Contributions

Taiwo Kolajo: conceptualization (equal), data curation (lead), formal analysis (lead), investigation (lead), methodology (equal), writing – original draft (lead), writing – review and editing (supporting). **Olawande Daramola:** conceptualization (equal), funding acquisition (lead), investigation (equal), methodology (equal), project administration (lead), resources (lead), supervision (lead), validation (lead), writing – original draft (supporting), writing – review and editing (lead).

Acknowledgments

The work is supported by the National Research Foundation (NRF), South Africa, University of Pretoria, South Africa, and Federal University Lokoja, Nigeria.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Related Wires Articles

Data stream mining in ubiquitous environments: State-of-the-art and current directions

Sentiment analysis for mining texts and social networks data: Methods and tools

Explainable artificial intelligence and machine learning: A reality rooted perspective

Combating disinformation in a social media age

Explainable artificial intelligence: an analytical review

References

Aboucaya, W., S. Guehis, and R. Angarita. 2023. "Building Online Public Consultation Knowledge Graphs." In *Text2KG 2023: International Workshops on Knowledge Graph Generation From Text Co-Located With ESWC 2023*, 12. Hersonissos (Crete), Greece.

Ai, L., J. Langer, S. H. Muggleton, and U. Schmid. 2023. "Explanatory Machine Learning for Sequential Human Teaching." *Machine Learning* 112: 3591–3632. https://doi.org/10.1007/s10994-023-06351-8.

Ali, S., T. Abuhmed, S. El-Sappagh, et al. 2023. "Explainable Artificial Intelligence (XAI): What Is Left to Attain Trustworthy Artificial Intelligence." *Information Fusion* 99: 101805. https://doi.org/10.1016/j.inffus.2023.101805.

Arsovski, S., H. Osipyan, M. I. Oladele, and A. D. Cheok, 2019. "Automatic Knowledge Extraction of Any Chatbot from Conversation." *Expert Systems With Applications* 137: 343–348. https://doi.org/10.1016/j.eswa.2019.07.014.

Arrieta, A. B., N. Diaz-Rodriguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

Bingley, W. J., C. Curtis, S. Lockey, et al. 2023. "Where Is the Human in Human-Centred AI? Insights From Developer Priorities and User Experiences." *Computers in Human Behaviour* 141: 107617. https://doi.org/10.1016/j.chb.2022.107617.

Borrego-Diaz, J., and J. G. Paez. 2022. "Knowledge Representation for Explainable Artificial Intelligence." *Complex & Intelligent Systems* 8: 1579–1601. https://doi.org/10.1007/s40747-021-00613-5.

Capua, M., E. Nardo, and A. Petrosino. 2016. "Unsupervised Cyberbullying Detection in Social Networks." In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 432–437. Cancun, Mexico.

Cardinale, Y., G. Freites, E. Valderrama, A. Aguilera, and C. Angsuchotmetee. 2022. "Semantic Framework of Event Detection in Emergency Situations for Smart Buildings." *Digital Communications and Networks* 8, no. 1: 64–79. https://doi.org/10.1016/j.dcan.2021.06.005.

Chakman, K., S. D. Swamy, A. Das, and S. Debbarma. 2020. "5W1H-Based Semantic Segmentation of Tweets for Event Detection Using BERT." In *Machine Learning, Image Processing, Network Security and Data Sciences, 1240, Communications in Computer and Information Science*, edited by A. Bhattacharjee, S. Borgohain, B. Soni, G. Verma, and X. Z. Gao, 57–72. Singapore: Springer. https://doi.org/10.1007/978-981-15-6315-7_5.

Chen, X., and Q. Li. 2020. "Event Modeling and Mining: A Long Journey Toward Explainable Events." *VLDB Journal* 29: 459–482. https://doi.org/10.1007/s00778-019-00545-0.

Chi, H., Z. Lin, H. Jin, X. Baoguang, and M. Qi. 2017. "A Decision Support System for Detecting Serial Crimes." *Knowledge-Based Systems* 123: 88–101. https://doi.org/10.1016/j.knosys.2017.02.017.

Choudhary, M., S. S. Chouhan, E. S. Pilli, and S. K. Vipparthi, 2021. "Berconvonet: A Deep Learning Framework for Fake News Classification." *Applied Soft Computing* 110: 107614. https://doi.org/10.1016/j.asoc.2021.107614.

Donadello, I., and M. Dragoni. 2021. "SeXAI: A Semantic Explainable Artificial Intelligence Framework." In *AIxIA 2020—Advances in Artificial Intelligence, 12414, Lecture Notes in Computer Science*, edited by M. Baldoni and S. Bandini. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-77091-4_4.

Doran, D., S. Schulz, and T. R. Besold. 2017. "What Does Explainable AI Really Mean? A New Conceptualisation of Perspectives." In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 Co-Located With 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA2017)*, edited by T. R. Besold and O. Kutz. Bari, Italy.

Drury, B., S. M. Drury, M. A. Rahman, and I. Ullah. 2022. "A Social Network of Crime: A Review of the Use of Social Network for Crime and the Detection of Crime." *Online Social Networks and Media* 30: 100211.

Egger, R. 2022. "Text Representations and Word Embeddings." In *Applied Data Science in Tourism.Tourism on the Verge*, edited by R. Egger. Cham, Switzerland: Springer.

Ehsan, U., K. Saha, M. D. Choudhury, and M. O. Riedl. 2023. "Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI." *Proceedings of the ACM on Human-Computer Interaction* 7, no. CSCW1: 34. https://doi.org/10.1145/3579467.

Gadek, G., and P. Guelorget. 2020. "An Interpretable Model to Measure Fakeness and Emotion in News." *Procedia Computer Science* 176: 78–87.

Garibay, O. O., B. Winslow, S. Andolina, et al. 2023. "Six Human-Centred Artificial Intelligence Grand Challenges." *International Journal of Human-Computer Interaction* 39, no. 3: 391–437.

Ghosal, T., T. Shaikh, T. Biswas, A. Ekbal, and P. Bhattacharyya. 2022. "Novelty Detection: A Perspective From Natural Language Processing." *Computational Linguistics* 48, no. 1: 77–117. https://doi.org/10.1162/coli_a_00429.

Goldani, M. H., S. Momtazi, and R. Safabakhsh. 2021. "Detecting Fake News With Capsule Neural Networks." *Applied Soft Computing* 101: 106991. https://doi.org/10.1016/j.asoc.2020.106991.

Hind, M. 2019. "Explaining Explainable AI. XRDS: Crossroads." *ACM Magazine for Students* 25, no. 3: 16–19. https://doi.org/10.1145/3325198.

Holzinger, A., A. Saranti, C. Molnar, P. Biecek, and W. Samek. 2022. "Explainable AI Methods—A Brief Overview." In *xxAI—Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science*, edited by A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Muller, and W. Samek, vol. 13200. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-04083-2_2.

Hu, X., W. Ma, C. Chen, et al. 2022. "Event Detection in Online Social Network: Methodologies, State-of-the-Art, and Evolution." *Computer Science Review* 46: 100500.

Ibrohim, M. O., and I. Budi. 2018. "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media." *Procedia Computer Science* 135: 222–229.

Kaur, S., S. Singh, and S. Kaushal. 2021. "Abusive Content Detection in Online Generation Data: A Survey." *Procedure Computer Science* 189: 274–281.

Khan, I., K. Ahmad, N. Gul, T. Khan, N. Ahmad, and A. Al-Fuqaha. 2023. "Explainable event recognition." *Multimedia Tools and Applications* 82: 40531–40557. https://doi.org/10.1007/s11042-023-14832-0.

Kocher, G., and G. Kumar. 2021. "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: Recent Developments and Challenges." *Soft Computing* 25: 9731–9763.

Kolajo, T., and O. Daramola. 2023. "Human-Centric and Semantics-Based Explainable Event Detection: A Survey." *Artificial Intelligence Review* 56: 119–158. https://doi.org/10.1007/s10462-023-10525-0.

Kolajo, T., O. Daramola, A. Adebiyi, and A. Seth. 2020. "A Framework for Preprocessing of Social Media Feeds Based on Integrated Local Knowledge Base." *Information Processing & Management* 57, no. 6: 102348.

Kolajo, T., O. Daramola, and A. A. Adebiyi. 2022. "Real-Time Event Detection in Social Media Streams Through Semantic Analysis of Noisy Terms." *Journal of Big Data* 9: 90. https://doi.org/10.1186/s40537-022-00642-y.

Kumar, A., V. Tyagi, and S. Das. 2021. "Deep Learning for Hate Speech Detection in Social Media." In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, 1–4. Kuala Lumpur, Malaysia.

Kumar, M., P. K. Singh, M. K. Maurya, and A. Shivhare. 2023. "A Survey on Event Detection Approaches for Sensor-Based IoT." *Internet of Things* 22: 100720. https://doi.org/10.1016/j.iot.2023.100720.

Lecue, F. 2020. "On the Role of Knowledge Graphs in Explainable AI." *Semantic Web* 11, no. 1: 41–51. https://doi.org/10.3233/SW-190374.

Lundberg, S. M., and S. I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 4765–4774. Long Beach, CA.

Ma, T., H. Zhou, Y. Tian, and N. Al-Nabhan. 2021. "A Rumor Detection Algorithm Based on Entity Recognition, Sentence Reconfiguration, and Ordinary Differential Equation Network." *Neurocomputing* 447: 224–234.

Macha, D., M. Kozielski, L. Wrobel, and M. Sikora. 2022. "RuleXAI—A Package for Rule-Based Explanations of Machine Learning Model." *SoftwareX* 20: 101209. https://doi.org/10.1016/j.softx.2022.101209.

Machlev, R., L. Heistrene, M. Perl, et al. 2022. "Explainable Artificial Intelligence (XAI) Techniques for Energy and Power Systems: Review, Challenges and Opportunities." *Energy and AI* 9: 100169.

Malviya, A., and R. K. Dwivedi. 2022. "Deceptive News Prediction Using Machine Learning Techniques." In *Proceedings of the ICISS 2022 on Intelligent Sustainable Systems. Lecture Notes in Networks and Systems*, edited by J. S. Raj, Y. Shi, D. Pelusi, and V. E. Balas, vol. 458, 355–367. Singapore: Springer.

Meel, P., and D. K. Vishwakarma. 2021. "A Temporal Ensembling Based Semi-Supervised ConvNet for the Detection of Fake News Article." *Expert Systems with Applications* 177: 115002. https://doi.org/10.1016/j.eswa.2021.115002.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights From the Social Sciences." *Artificial Intelligence* 267: 1–38.

Mishima, K., and N. H. Yamana. 2022. "A Survey on Explainable Fake News Detection." *IEICE Transactions on Information and Systems* E105-D, no. 7: 1249–1257.

Moran, S., R. McCreadie, C. Macdonald, and L. Ounis. 2016. *Enhancing First Story Detection Using Word Embedding. SIGIR'16*, 821–824. Pisa, Italy: ACM.

Nahar, V., S. Al-Maskari, X. Li, and C. Pang. 2014. "Semi-Supervised Learning for Cyberbullying Detection in Social Networks." In *Databases Theory and Applications, 8506. Lecture Notes in Computer Science*, edited by H. Wang and M. A. Sharaf, 160–171. Cham, Switzerland: Springer.

Osborne, M., S. Petrovic, R. McCreadie, C. Macdonald, and L. Ounis. 2012. *Bieber no More: First Story Detection Using Twitter and Wikipedia. TAIA'12*, 1–4. Portland, OR: ACM.

Panagiotou, N., C. Akkaya, K. Tsioutsiouliklis, V. Kalogeraki, and D. Gunopulos. 2016. "First Story Detection Using Entities and Relations." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3237–3244. Osaka, Japan.

Panagiotou, N., C. Akkaya, K. Tsioutsiouliklis, V. Kalogeraki, and D. Gunopulos. 2021. "A General Framework for First Story Detection Utilising Entities and Their Relations." *IEEE Transactions on Knowledge and Data Engineering* 33, no. 11: 3482–3493. https://doi.org/10.1109/TKDE.2020.2970051.

Pesquita, C. 2021. "Towards Semantic Integration for Explainable Artificial Intelligence in Biomedical Domain." In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021)*, 747–753. Vienna, Austria. https://doi.org/10.5220/0010389707470753.

Qiu, Y., S. Li, R. Li, L. Wang, and B. Wang. 2015. "Nugget-Based First Story Detection in Twitter Stream." In *Social Media Processing, 568. Communications in Computer and Information Science*, edited by X. Zhang, M. Sun, Z. Wang, and X. Huang, 74–82. Singapore: Springer.

Qiu, Y., S. Li, W. Yang, R. Li, L. Wang, and B. Wang. 2016. "Time-Aware First Story Detection in Twitter Stream." In *2016 IEEE First International Conference on Data Science in Cyberspace*, 608–613. Changsha, China.

Rony, S., S. C. Bakchy, and H. Rahman. 2020. "Crime detection using data mining techniques." *Computer Science & Engineering* 10, no. 5: 1–5.

Rosa, H., N. Pereira, R. Ribeiro, et al. 2019. "Automatic Cyberbullying Detection: A Systematic Review." *Computers in Human Behaviour* 93: 333–345.

Seerat, B., and U. Qamar. 2015. "Rule Induction Using Enhanced RIPPER Algorithm for Clinical Decision Support System." In *2015 Sixth International Conference on Intelligent Control and Information Processing (ICICIP)*, 83–91. Wuhan, China. https://doi.org/10.1109/ICICIP.2015.7388149.

Sheikhi, S. 2021. "An Effective Fake News Detection Method Using WOA-xgbTree Algorithm and Content-Based Features." *Applied Soft Computing* 109: 107559. https://doi.org/10.1016/j.asoc.2021.107559.

Tiddi, I., and S. Schlobach. 2022. "Knowledge Graphs as Tools for Explainable Machine Learning: A Survey." *Artificial Intelligence* 302: 103627.

Trueman, T. E., K. J. Ashok, P. Narayanasamy, and J. Vidya. 2021. "Attention-Based C-BiLSTM for Fake News Detection." *Applied Soft Computing* 110: 107600. https://doi.org/10.1016/j.asoc.2021.107600.

Vemula, S. 2022. "*Human-Centred Explainable Artificial Intelligence for Anomaly Detection in Quality Inspection: A Collaborative Approach to Bridge the Gap Between Human and AI*." PhD diss., University of Incarnate Word.

Vuurens, J. B. P. A., and P. de Vries. 2016. "First Story Detection Using Multiple Nearest Neighbors." In *SIGIR'16*, 845–848. Pisa, Italy.

Wang, F. 2019. "*Distance, Time and Terms in First Story Detection*." PhD diss., Department of Comp. Infor. Sci., Technological University Dublin. https://doi.org/10.21427/spp0-zx14.

Wang, H., F. Shakerin, and G. Gupta. 2022. "FOLD-RM: A Scalable, Efficient, and Explainable Inductive Learning Algorithm for Multi-Category Classification of Mixed Data." *Theory and Practice of Logic Programming* 22, no. 5: 658–677. https://doi.org/10.1017/S1471068422000205.

Wurzer, D., and Y. Qin. 2018. "Parameterising kterm hashing." In *SIGIR'18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 945–948. Ann Arbor, MI.

Xiang, G., J. Hong, and C. P. Rosé. 2012. "Detecting Offensive Tweets via Topical Feature Discovery Over a Large-Scale Twitter Corpus." In *Proceedings of the 21st ACM Conference on Information and Knowledge Management*, 1980–1984. Maui, HI.

Yuan, H., H. Yu, S. Gui, and S. Ji. 2023. "Explainability in Graph Neural Networks: A Taxonomic Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45: 5782–5799. https://doi.org/10.1109/TPAMI.2022.3204236.

Zhao, R., A. Zhou, and K. Mao. 2016. "Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features." In *Proceedings of the 17th International Conference on Distributed Computing and Networking*, 1–6. New York, NY.