https://doi.org/10.1186/s40537-024-01040-2

Journal of Big Data

RESEARCH Open Access

Watch and learn: event-domain term extraction from social networks

Nicholas Mamo[1*], Joel Azzopardi[1] and Colin Layfield[1]

*Correspondence: nicholas.mamo.14@um.edu.mt
[1] Faculty of ICT, University of Malta, Msida MSD 2080, Malta

Abstract

Event tracking algorithms detect and track, but they do not understand what happens in events. Term extraction research has studied the concepts of general domains-computer science, medicine, law-but not What happens in event domains, and not from noisy social networks, where they are popularly narrated. The event structure, the message and its form distinguish event domains from general domains, and formal text from user-generated content. In this article, we present the Event-Aware Term Extractor (EVATE), the first term extractor built for event domains, and the first built for user-generated content. EVATE learns semantically: it tracks events to extract terms that describe What happens, and then ranks them with a termhood statistic designed for event domains. We compared our novel approach with four traditional term extractors in three disparate event domains on data from Twitter (now X). Because EVATE learns semantically, its lexicons described What happens in events better than standard approaches. Even when the term extractors could not adapt to unorthodox event domains, our novel method propped up the others as a semantic re-ranker. The results show that we need algorithms designed for event domains and for user-generated content. Crucially, they also show that we only need one semantic extractor like EVATE to adapt traditional algorithms.

Introduction

What makes a word a domain term? Is it its frequency in a domain, the specificity of its meaning, or some deeper semantic property? Research in Automatic Term Extraction (ATE), the task of identifying domain-specific terms from domain-specific corpora, has provided various answers to the question, but it has given them only for general domains and for formal texts. It has overlooked altogether event domains and user-generated content.

As for event-related research, it too has made few and futile efforts to understand What happens in event domains. Event tracking algorithms detect and track news events in microblogs, news reports and other documents, but they detect and track without understanding. Early research did attempt to understand which keywords drive the narrative, but it understood linguistically, by identifying and boosting nouns and verbs, to limited success [1]. Likewise, literature has made few efforts to extract terms from user-generated content, which became the *de facto* standard medium for event tracking [2]. Event tracking literature cast aside understanding [3, 4].

In this article, we understand event domains from user-generated content. We argue that event tracking needs tailored understanding, and hence tailored solutions for term extraction. We present the Event-Aware Term Extractor (EVATE), the first term extractor designed for event domains and for user-generated content. EVATE learns about event domains semantically, by 'watching' events: it uses an event tracking algorithm to extract topical keywords from events, then ranks them with a novel termhood measure. We make the following contributions:

- We present EVATE, the first term extractor designed for event domains and for user-generated content. Unlike traditional term extractors, which approximate semantics through linguistic and statistical analysis, our method injects semantics by using an event tracker to extract topical terms that describe *What* happens in events.
- We apply EVATE and, for the first time, four traditional term extractors in three disparate event domains and on user-generated content. Our findings confirm that the event structure and informal text pose vastly different challenges to term extraction than general domains and formal texts.
- We use EVATE as a semantic re-ranker to adapt the four traditional term extractors to event domains. The final lexicons prove that a semantic term extractor like EVATE can adapt traditional assumptions to event domains and to user-generated content.

The rest of this article is structured as follows. In Sect. 2, we discuss the distinctive challenges of event domains and user-generated content. Then, in Sect. 3, we propose EVATE, the first term extractor designed to overcome those challenges. In Sect. 4, we compare EVATE with four traditional term extractors on tweets from three event domains. We conclude the article with directions for future work in Sect. 5.

Related work

The contemporary term extractor has two components: a linguistic component and a statistical one [5]. The linguistic component uses a syntactical filter to extract candidates: nouns, mostly, and sometimes verbs [6]. The linguistic component exists only to compensate for the simplistic assumptions of its statistical counterpart [7], which estimates termhood from rough indicators such as frequency, specificity and consistency. Despite the assumptions, the contemporary term extractor performs remarkably well.

Rather, we know that the contemporary term extractor performs remarkably well on the formal texts of general domains: computer science [8], medicine [7], law [9]. We could find little research that could stand for term extraction on user-generated content from event domains. What we found relied on manual annotation [10, 11] or feature-selection techniques [12], or extracted terms from news reports instead of tweets [13, 14]. Neither evaluated the terms themselves-only their effects.

It would be too bold to assume that what works for formal text will work for user-generated content, and that what works in general domains will work in event domains. On the contrary, in this section we identify several consequential differences. We start by naming three principal differences between formal texts and user-generated content.

First, term extraction from user-generated content must adapt to inferior writing quality. Social network users write in a "non-standard way" [12]. They reject the formal language of medical and legal texts for a more informal personality that thrives in brevity, and accepts abbreviations and mistakes [15]. How well do the linguistic components' syntactical choices navigate noise, and the statistical components' estimates adjust to brevity?

Second, term extraction must adapt to new types of subject matter. A news report becomes newsworthy the moment a newsroom publishes it, but not tweets, and not their noise [13]. Noise takes an obvious form in advertisement and spam. Meladianos et al. [16] found noise to have increased drastically since 2012, and more recently, Hasan et al. [17]'s hand-picked list of around 350 spam phrases filtered 70% of tweets. More ordinarily, however, noise appears simply as the absence of relevance: the everyday and opinionated conversations, which make up nearly half of all tweets [13].

Third, term extraction must adapt to repetition. Repetition too takes an obvious form in Twitter's redundant retweets, Tumblr's re-blogs and Mastodon's boosts, which contribute no new information and accentuate bias [18]. On Twitter (now X), McMinn and Jose [19] found retweets to account for nearly 30% of all tweets. Repetition, however, also takes a more insidious form: the same event narrated from the individual perspectives of a myriad users. Such repetition reveals the popularity of events, but popularity does not always translate into relevance or importance. The contemporary term extractor must know to recognise and handle all the different forms of noise in modern media.

Even if the differences between mediums were immaterial, we cannot assume that what works in general domains also works in events and event domains. We define events and event domains as follows:

Definition 1 (Event) "An action, or a series of actions, or a change [What] that happens at [a] specific time [When] due to specific reasons [How/Why], with associated entities such as objects, humans [Who], and locations [Where]." [20]

Definition 2 (Event domain) A prototypical representation of a set of related events with a common structure: Who usually does What, Where and When, and Why and How.

The structure of events-Who does What, Where and When, and Why and How-informs the design of our event domain term extractors. In particular, we identify four considerations of event domains that do not exist in general domains.

First, the structure itself shifts across and within event domains. Some domains give a prominent role to certain aspects of the structure; in localised events, like football matches, the location Where the event happens does not change, unlike What happens or Who makes it happen. Other domains may have an undefined structure; in general breaking news detection, we can neither anticipate the events [15, 21] nor, consequently, their structure. Domains may change over time too, as do current affairs in politics. The assumptions that work in one domain may fail in another.

Second, named entities play a bigger role in event domains than in general domains. Consider Chung [22]'s specificity scale, adapted to the event domain of football matches in Fig. 1. Chung [22] assumed that the more specific a term, the higher its termhood. Without any linguistic filters, the most specific terms of event domains describe named entities; *Lionel Messi* and *Chelsea* are relevant almost exclusively in football. In fact, most named entities describe only a few events, not the entire event domain-event terms, not domain terms. Event-domain term extractors must distinguish between event terms and domain terms.

Third, the semantics of terms differ between general domains and event domains. No terms define event domains quite like the words that describe What happens within: goals in football matches, overtakes in Grands Prix, votes in elections. When early event tracking research needed to understand What happens in events, it understood through nouns and verbs [23, 24]. Such approximative understanding barely improved results [1], and only after researchers filtered it semantically [24]. As a result, even today, understanding rarely drives event tracking [25].

Previously, we argued that those limited improvements do not reflect poorly on the idea of understanding [3, 4]. Rather, they reflect poorly on the interpretation of understanding: what transforms a noun or a verb into a term in event domains. Nouns and verbs can be helpful, but left unchecked, they introduce noise [26]. In event domains, algorithms do not need any nouns and verbs, but only those that capture the actions and changes-What happens in event domains [20].

Fourth, many event domains render the ideal corpora of general-domain term extraction unrealistic. The ideal dataset is thematic, accurate and randomised [27], and in event domains, the ideal dataset often lies out of reach. Countries only hold a national election every few years, and in Formula 1, some drivers and constructors have competed almost every single year since Twitter's launch. In many event domains, it would be impossible to curate a balanced and randomised sample of events from Twitter, let alone from newer social networks.

None of the surveyed algorithms address the realities of term extraction from user-generated content or event domains. To extract terms from user-generated content

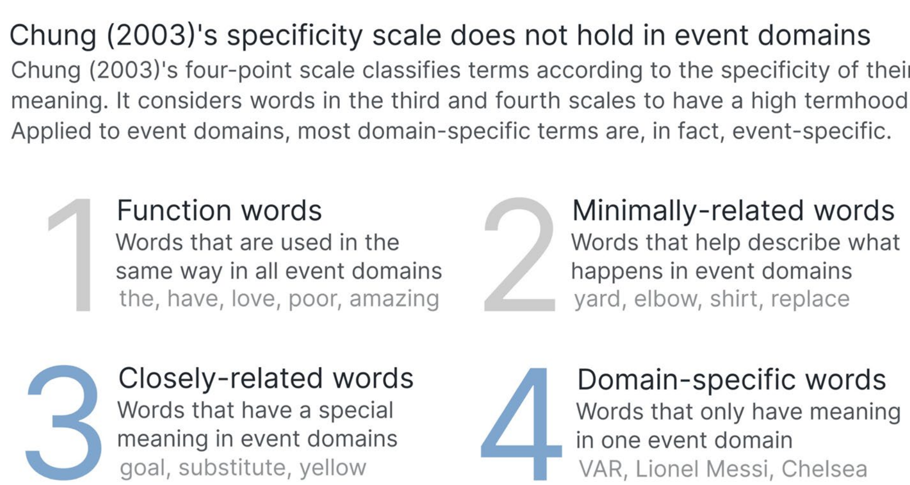

Fig. 1 Chung [22]'s specificity scale does not hold in event domains. Domain-specific words, which would denote model terms in more traditional domains, describe such a small subset of the domain that they become event-specific

and from event domains, the contemporary term extractor must adopt new principles. It must understand how social networks function and how their users converse, and it must understand, semantically, What happens in events. In the next section, we propose EVATE, a term extractor built on those new principles.

Methodology

A common thread connects term extraction, whether on formal content or user-generated content, or whether in general domains or event domains: semantics. Termhood should measure "semantic informativeness" [28]. When we say that the interpretation of understanding failed early event tracking research, we mean that it did not capture semantic informativeness. The Event-Aware Term Extractor (EVATE), summarized in Fig. 2, captures our interpretation of semantic event understanding. Unlike the contemporary term extractor, EVATE's linguistic and statistical components share the role of extracting semantically-meaningful terms. We explain how in the rest of this section.

The linguistic component

A simple intuition drives EVATE. We assume that since event domain terms describe What happens in events, they describe sub-events. If *speech* represents a political domain term, then we will observe speech sub-events. Therefore, as its linguistic component, EVATE uses an event tracking algorithm to track events, identifying sub-events and, crucially, their topical keywords. Eventually, the topical keywords become the domain terms, which the statistical component ranks. We define sub-events and topical keywords as follows:

Definition 3 (Sub-event) An event within an event, with an identical structure: Who does What, Where and When, and Why and How.

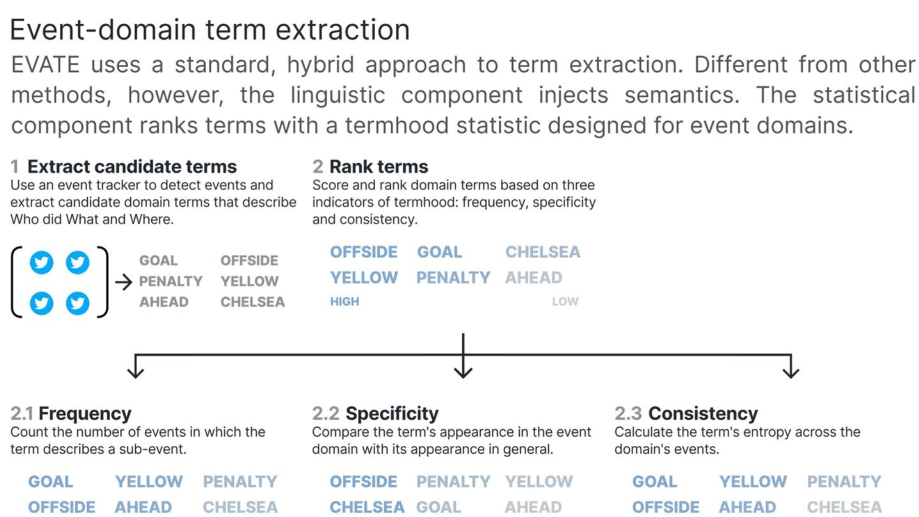

Fig. 2 EVATE's architecture injects semantics through its linguistic component. A statistical measure ranks terms based on three qualities: frequency, specificity and consistency

Definition 4 (Topical keyword) Words or phrases that describe Who did What, Where and When, and Why and How.

As our linguistic component, we use Event TimeLine Detection (ELD) [29]. ELD alone overcomes many of the challenges from Sect. 2. In-built filters remove spam, advertisements and other types of noise to maximise precision. A clustering algorithm groups tweets into clusters to minimise repetition, and a statistical measure confirms whether clusters represent sub-events by looking for bursty topical keywords: EVATE's candidate domain terms. While any other event tracker that extracts topical keywords can substitute, EVATE mirrors the event tracker, in qualities as in flaws. ELD only extracts unigram topical keywords, and so does EVATE.

The statistical component

The linguistic component's topical keywords have a semantic bearing, some more than others. Therefore, EVATE's statistical component takes a less determining role, that of ranking topical keywords based on a termhood measure. The statistical component looks for three qualities: a term is relevant to the domain, carries a specific meaning, and appears consistently. We describe how we measure the three qualities in more detail next.

First, Event Frequency (EF) measures relevance. An event domain term should describe What happens in many of the domain's events. Therefore, EF counts the number of events in set E in which a term t emerges as a topical keyword. Here, $T(e)$ represents the set of topical keywords of event e. We measure EF as follows:

$$EF_t = \log |\{e \in E | t \in T(e)\}|. \tag{1}$$

We note two things about EF. First, EF counts the number of events, not the number of sub-events, because some types of sub-events can occur a limited number of times in an event; an electoral campaign has several speeches but only one victory. Second, we take the logarithm, both because a term only needs to appear a few times to be relevant [7] and to allow rare terms to climb the ranking.

Second, Inverse Corpus Frequency (ICF) measures specificity. An event domain term should appear more frequently in the event domain than in general. We base our formulation on Reed et al. [30]'s Term Frequency-Inverse Corpus Frequency (TF-ICF), which estimates the second component ofTerm Frequency-Inverse Document Frequency (TF-IDF) on a static corpus, G. We measure ICF as follows:

$$ICF_t = \log \frac{|G|}{|\{d \in G | t \in d\}| + 1}. \tag{2}$$

Our general corpus consists of 457,429 English tweets, which we collected over 12 h using the Twitter Sample API between 11 April and 12 April, 2020. ICF thus punishes noisy or everyday words, like *stream*, *love* or *best*.

Third, entropy measures consistency. An event domain term should be relevant even when it does not emerge as topical. We use entropy to prioritise terms that distribute uniformly across the domain's events, where $D(e)$ represents the set of documents of event e. We measure entropy as follows:

$$p_{t,e} = \frac{|\{d \in D(e)|t \in d\}|}{\sum_{e'}^{E} |\{d \in D(e')|t \in d\}|}. \tag{3}$$

$$Entropy_t = -\sum_{e \in E} p_{t,e} \log p_{t,e}. \tag{4}$$

Consistency matters because as we showed in Fig. 1, the most specific terms tend to represent event terms, not domain terms. The difference between domain terms and event terms is that the former are not merely relevant but consistently so; even when a football match has no goals, users will talk about *goals*. Entropy punishes the relevant and highly-specific but still-inconsistent event terms.

EVATE's final termhood measure combines relevance, specificity and consistency. An event domain term exhibits all three qualities at once, so we multiply EF, ICF and entropy:

$$EVATE_t = EF_t \cdot ICF_t \cdot Entropy_t. \tag{5}$$

Note that the logarithmic bases in EF and entropy scale the termhood scores but do not affect the rankings. We evaluate EVATE in three event domains next.

Evaluation

In this section, we evaluate EVATE in three event domains: football matches, Formula 1 Grands Prix and politics. While we evaluate on tweets, the preferred medium of event tracking [2], we believe that our conclusions apply to other types of user-generated content too. In the absence of term extractors built for event domains and for user-generated content, we compare EVATE with four general-domain term extractors:

- Term Frequency-Disjoint Corpora Frequency (TF-DCF), which accepts that a domain term may appear in a few different domains [7]
- Term Frequency-Inverse Corpus Frequency (TF-ICF), which adapts TF-IDF by calculating IDF on a static corpus [30]. We use TF-ICF since TF-IDF would assign a low weight to terms that appear in many domain-specific documents [31].
- Domain Specificity, which prioritises terms that appear predominantly in the event domain more than in general [32]
- Rank Difference, which favours terms that rank higher in the event domain than in general [33]. We rank terms using frequency, but due to the large number of out-of-dictionary words in user-generated content, we set a minimum frequency of 100 when working with all tweets, and 50 when working with tweets by verified users.

The four baselines share a common linguistic component: stemmed nouns, verbs and adjectives, which we extract using NLTK [34]. They also share EVATE's general corpus, which their statistical components use to contrast the appearance of a term in the domain with its appearance in general. We describe our experimental set-up and results in the rest of this section.

The language of football matches

The event domain of football matches comes closest to the ideal datasets of term extraction research. Football matches have an almost-identical structure, well-defined rules and clear boundaries. Therefore, we base our first analyses on a varied set of 24 football matches from the period between 16 May and 21 August, 2020. In total, we collected 4,618,221 tweets, including an hour-long understanding period from before each match; from them, ELD constructs a TF-ICF term-weighting scheme [29]. We evaluate against a dictionary by UEFA [35] and a crowd-sourced glossary from Wikipedia [36].

The four baselines struggled to adapt to user-generated content. Their results contrasted sharply with those of term extraction on formal texts. Still, as Table 1 shows, results improved when we only extracted terms from tweets by verified authors. When we collected the datasets, Twitter still verified users selectively: celebrities, journalists and brands. Authoritative users write authoritative content, and authoritative content resembles the formal text of traditional term extraction. There, in their element, the traditional extractors performed better. Term extraction on user-generated content requires data curation, if not tailored techniques.

In reality, the results look worse than they actually are. The glossaries include many multi-word terms, incompatible with our mono-word terms: *yellow card*, not *yellow* and *card* separately. The ground truth lists also omit new terms, like *VAR*, and favour technical terminology: *offside position*, not *offside*. Similarly, they exclude informal speech, a staple of user-generated content: *substitution*, not *sub*; *penalty*, not *pen*. Term extraction on user-generated content also needs ground truths that reflect social media discourse.

The baselines also struggled to adapt to event domains. TF-ICF and TF-DCF, which prioritised frequent terms, chose general words-not necessarily wrong but superficial. TF-ICF's first four terms-*goal*, *score*, *player* and *game*-describe several sports, not just football. Conversely, Rank Difference and Domain Specificity pushed specific event terms, like named entities and hashtags, which NLTK occasionally mistook for nouns. Neither extracted the kind of understanding that event tracking research needs.

Here, EVATE stood out with two qualities. First, EVATE captured What happens in events better than the baselines. We evaluated again on two smaller ground truths compiled specifically to summarise What happens in football matches [37, 38].

Table 1 Traditional term extractors require careful data curation but still do not capture What happens in event domains

	P@50		P@100		P@200	
	All (%)	Verified (%)	All (%)	Verified (%)	All (%)	Verified (%)
EVATE	50.00		36.00		29.00	
EVATE$_{\text{bootstrapped}}$	**54.00**		**45.00**		29.00	
TF-ICF	46.00	48.00	34.00	43.00	26.00	31.50
TF-DCF	38.00	46.00	29.00	38.00	24.50	30.00
Domain Specificity	22.00	32.00	21.00	25.00	19.50	22.50
Rank Difference	32.00	36.00	32.00	37.00	25.00	**42.50**

Best results in bold

EVATE obtained the highest recall (52.50%), ahead of TF-ICF (50.00%) and Rank Difference (47.50%). Term extraction in event domains needs ground truths that capture What happens in events.

Second, EVATE had a high precision at the top. No baseline achieved a higher *P*@50, and only one term among the top ten was clearly irrelevant: *FFS*. The high precision at the top permitted further processing to further emphasise EVATE's semantic properties: bootstrapping.

We followed a simple bootstrapping procedure. Using the top *k* terms as the seed set, we applied the chi-square statistic to bootstrap the remaining 200-*k* terms. We calculated the statistic based on whether two terms appeared in the same tweet, excluding retweets. At every iteration, we bootstrapped the ten terms having the highest average chi-square score with seed terms and previously-bootstrapped terms.

Bootstrapping performed best with 16 seed terms. Semantically-meaningful terms bootstrapped other semantically-meaningful terms. Terms like *corner*, *net* and [referee] *decision* rose over a hundred ranks ahead of subjective expressions and profanity, like *masterclass*, *wow* and *WTF*. EVATE's *P*@50, already the highest, rose from 50% to 54%, and *P*@100 from 36% to 45%-a new high. Despite the incomplete ground truths, precision only dropped below 50% after 88 terms. Thus, EVATE built a precise lexicon with the kind of semantic understanding that event tracking requires.

The language of Formula 1 Grands Prix

The event domain of Formula 1 Grands Prix changes too slowly to collect the ideal datasets of term extraction research. Drivers and constructors, Who race in Grands Prix, change over years, not weeks, and blur the separation between event terms and domain terms. We base our analyses on 15 Grands Prix from the 2020 season. In total, we collected 2,208,477 tweets, including half-hour-long understanding periods, but following our previous analyses, we only used tweets by verified authors. We evaluate against dictionaries by Formula 1 [39], F1technical [40] and Formula 1 Dictionary [41], and a crowd-sourced glossary from Wikipedia [42]. Due to the technical nature of the event domain, we split multi-word terms into mono-word terms.

All algorithms struggled with event terms. TF-ICF and TF-DCF struggled because users mentioned locations as adjectives: *Austrian* or *Styrian* Grand Prix. Rank Difference and Domain Specificity struggled because NLTK sometimes mistook drivers and constructors for common nouns. Similarly in EVATE, drivers and constructors emerged as topical regularly, and occupied 17 out of the top 20 ranks. Consequently, as Table 2 shows, our method obtained a low precision throughout. Neither algorithm produced accurate, semantic understanding.

Nevertheless, the baselines and EVATE complement each other. The baseline's linguistic component filters named entities; EVATE's injects semantics. Likewise, the baselines' statistical components promote general or specific terms; EVATE's penalises event terms. Therefore, we employed our method as a semantic re-ranker. Each baseline generated a lexicon of terms, for which EVATE calculated its own termhood scores. Then, we rescaled the baseline's and EVATE's scores separately between 0 and 1, and multiplied them to get the final termhood score.

Table 2 As a re-ranker, EVATE lent its qualities-semantics and high precision-to the traditional term extractors

	P@50		P@100		P@200	
	Base (%)	Re-ranked (%)	Base (%)	Re-ranked (%)	Base (%)	Re-ranked (%)
EVATE	24.00%		28.00		25.50	
TF-ICF	58.00	**66.00**	47.00	48.00	37.50	37.50
TF-DCF	48.00	56.00	47.00	48.00	37.50	37.50
Domain Specificity	30.00	42.00	23.00	27.00	18.50	18.50
Rank Difference	38.00	60.00	45.00	**49.00**	39.00%	**39.00**

Best results in bold

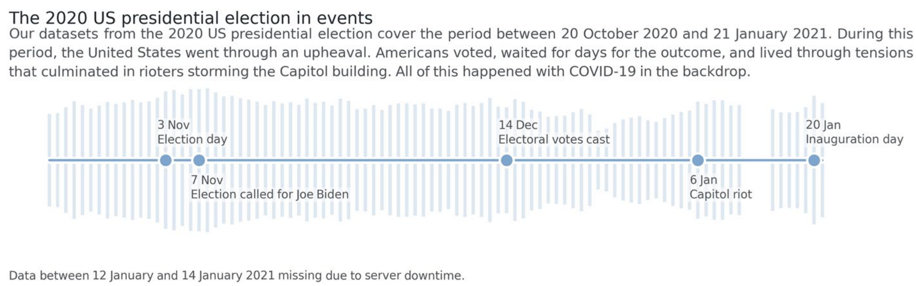

Fig. 3 The relative tweet volume in our political corpora. The data spans three months and covers, among other events, the 2020 US presidential election, the riots at the Capitol and inauguration day

EVATE's semantic re-ranking improved every baseline's lexicon. The improvements in *P*@50 ranged from 8% in TF-ICF and TF-DCF to 22% in Rank Difference. Rank Difference only obtained 6% lower *P*@50 than TF-ICF, which maintained two-thirds precision after 50 terms, and outperformed it in *P*@100. In TF-ICF, event terms and everyday words like *Austrian* (3 ▽ 115), *take* (19 ▽ 122) and [they]*'re* (23 ▽ 125) dropped more than a hundred ranks, their place taken by more semantic domain terms, such as penalty (42 △ 13), *point* (34 △ 15) and *retire* (47 △ 16). Every lexicon became more precise, more semantic.

More than just improving performance, EVATE adapted traditional algorithms to event domains. Despite its own difficulties, and those of traditional term extractors, EVATE as a re-ranker built semantically-meaningful lexicons. EVATE transformed TF-ICF and TF-DCF from algorithms that describe domains generally into algorithms that describe general sub-events, and Rank Difference and Domain Specificity from algorithms that describe domains technically into algorithms that describe technical sub-events. Moreover, by replacing the noise in the top ranks with semantically-meaningful terms, re-ranking allows bootstrapping. We combine re-ranking and bootstrapping next.

The language of politics

The event domain of politics has no structure: a myriad events with no clear start and no clear end, governed by no fixed rules. For term extraction to work, it must work with imperfect data. In the last experiments, we use EVATE to create a transferable and semantic political lexicon. We base our analyses on data from the 2020 US presidential election, between 20 October, 2020 and 21 January, 2021. In total, we collected

85,518,453 tweets, excluding retweets, by tracking several hashtags, and Donald Trump and Joe Biden. Figure 3 shows the tweet distribution and key events in our datasets.

We kept the same configurations as before, except EVATE's, for two reasons. First, EVATE expects every dataset to contain one event, infeasible in politics. We assumed simply that each day represents a separate event. Second, because we cannot predict when an event starts, we cannot establish ELD's understanding period and, consequently, the term-weighting scheme. Instead, we used the general corpus from Sect. 3, thus boosting words specific to politics. We evaluate against political terminology by the Dole Institute of Politics [43] and Brown et al. [44], legal terminology by Justia [45] and the United States Courts [46], and election terminology by the U.S. Election Assistance Commission [47].

To understand the domain's difficulties, consider Table 3. TF-ICF and TF-DCF built general lexicons: at the top, terms like *election*, *vote* and *president*. Yet the everyday talk between ephemeral, one-time events marred the lexicons; no other lexicons included as many of English First [48]'s 1000 most common words. Conversely, Domain Specificity and Rank Difference rejected common words-even correct terms-for a more technical extreme: at the top, terms like *coup*, *cabinet* and *runoff*. Finally, EVATE balanced the general with the technical, but without structure it filled the lexicon with event terms: political figures, states and counties.

We base our transferable and semantic political lexicon on TF-ICF's. While Rank Difference performs best, its lexicon shares little overlap with EVATE's-just 175 out of the top 1,000 terms, as opposed to 382 of TF-ICF's top 1000 terms. EVATE would have a meagre influence. Moreover, by its general nature, TF-ICF's lexicon has all the trappings of a transferable lexicon. It has only two flaws, both of which re-ranking can solve: excessive noise and lack of semantics.

As a re-ranker, EVATE sorted the top 1,000 words in TF-ICF's lexicon. As Table 4 shows, *P*@50 increased by 16%, *P*@100 by 7% and *P*@200 by 6%. The share of common words plummeted. Noisy, everyday words like *say* (4 $\triangledown$ 384), *think* (32 $\triangledown$ 389) and *want* (49 $\triangledown$ 392) gave way for semantically-meaningful terms: *county* (352 $\triangle$ 99), *riot* (247 $\triangle$ 96) and *senate* (209 $\triangle$ 62). Not only did EVATE reduce noise, but it also improved the order quality at the top.

It is the same quality at the top that permits bootstrapping. We bootstrapped the top 400 terms from the re-ranked lexicon, ten at a time. The best configuration needed

Table 3 In the event domain of politics, the lexicons were either inaccurate, noisy or overly-technical, and thus, not transferable

	Event domain terms			Common words [48]		
	P@50 (%)	P@100 (%)	P@200 (%)	P@50 (%)	P@100 (%)	P@200 (%)
EVATE	24.00	28.00	26.00	14.00	17.00	22.50
TF-ICF	**46.00**	**38.00%**	32.00	68.00	76.00	74.50
TF-DCF	44.00	**38.00**	30.50	74.00	73.00	72.50
Domain Specificity	30.00	27.00	28.00	**6.00**	**3.00**	**4.00**
Rank Difference	44.00	**38.00**	**35.50**	8.00	6.00	6.50

Best results in bold

Table 4 As a semantic re-ranker, EVATE cleaned TF-ICF's lexicon and made it more semantic, whereas bootstrapping further improved precision

	Event domain terms			Common words [48]		
	P@50 (%)	P@100 (%)	P@200 (%)	P@50 (%)	P@100 (%)	P@200 (%)
EVATE	24.00%	28.00	26.00	**14.00**	**17.00**	**22.50**
TF-ICF	46.00	38.00	32.00	68.00	76.00	74.50
TF-ICF + EVATE	62.00	45.00	38.00	56.00	58.00	62.50
TF-ICF + EVATE with bootstrapping	**68.00**	**57.00**	**42.00**	44.00	43.00	51.00

Best results in bold

just six seed terms. Of 109 precise terms, bootstrapping moved 57-more than half-to the top 100, and 84 to the top 200. Event terms like *Biden* (57 ▽ 380) and other general words dropped drastically. Instead rose event domain terms: *mail*[-in ballot] (359 △ 23), *cast* [ballot] (216 △ 13) and *project* (268 △ 66). Compared to TF-ICF's original lexicon, *P*@50 rose by 22%, *P*@100 by 19% and *P*@200 by 10%, and the share of common words dropped by 23.50%.

The final lexicon combines the best qualities of TF-ICF with the best qualities of EVATE. Evidently, it misses the mannerisms of unverified social network users, the technical terms of Rank Difference, and language from alternative systems of governance. Still, the final lexicon establishes a baseline of political discourse: the laws (*court*, *fraud* and *investigation*), the politics (*vote*, *campaign* and *veto*), the policies (*education*, *health* and *immigration*). Much of the credit comes back to EVATE: EVATE the re-ranker, and the enabler of bootstrapping. We conclude this article next.

Conclusion

Frequency and specificity, and nouns and verbs only approximate semantics. In this article, we showed that such approximations suffice for the formal content of general domains, but not for user-generated content, and not for event domains. However, we also showed that we only need one semantic extractor to adapt traditional assumptions. From sports to politics, EVATE adapted to different event structures and extracted semantically-meaningful domain terms. Evidently, as the first term extractor designed for event domains and user-generated content, some limitations hinder EVATE. We outline three avenues for future work:

- Multi-word term extraction. In this work, we only extracted single-word terms as we prioritised semantic informativeness, but phrases can also be terms [28]. Future work should extend EVATE to extract multi-word terms, such as by identifying single-word terms that can combine into longer expressions, like *yellow* and *card*.
- Conversational ground truths. In this work, we showed that people speak differently on social media than in formal texts, but existing ground truths do not cater to those differences. Future work should create new ground truths that capture the linguistic nuances of user-generated content.
- Event-domain term extraction from formal content. Restrictive social media policies are making it increasingly difficult to access user-generated content, but we can still

understand event domains from formal content. Future work should explore whether keyphrase extraction from news articles and live blogs can replace EVATE's linguistic component.

Above all, future work should return to the question that animated early event tracking research, whether understanding can improve event tracking. We reiterate our hypothesis [3, 4]: it was not the idea of understanding that failed event tracking research but the interpretation of understanding. EVATE's terms give a new, more semantic interpretation of understanding. Future work should explore understanding's role in event tracking anew, to give an answer that only semantically-meaningful terms like EVATE's can provide.

Acknowledgements
Not applicable.

Author contributions
Conceptualisation: N.M., J.A. and C.L.; Methodology: N.M., J.A. and C.L.; Software: N.M.; Validation, N.M.; Formal Analysis: N.M.; Investigation: N.M.; Resources: N.M., J.A. and C.L.; Data Curation: N.M.; Writing—Original Draft Preparation: N.M.; Writing—Review & Editing: N.M., J.A. and C.L.; Supervision: J.A. and C.L.; Project Administration: J.A. and C.L.; Funding Acquisition: N.M. and J.A. All authors read and approved the final manuscript.

Funding
This research work disclosed in this publication was partially funded by the Tertiary Education Scholarships Scheme (Malta).

Data availability
The datasets generated and/or analysed during the current study are available in the `NicholasMamo/EVATE-data` repository, https://github.com/NicholasMamo/EVATE-data. The algorithms developed and used in the current study are available in the `NicholasMamo/EvenTDT` repository, https://github.com/NicholasMamo/EvenTDT.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 February 2024 Accepted: 14 December 2024
Published online: 26 December 2024

References

1. Mohd M. Named entity patterns across news domains. In: Proceedings of the BCS IRSG Symposium: Future Directions in Information Access 2007. Glasgow, Scotland: BCS, The Chartered Institute for IT; 2007;1–6. Available from: https://dl.acm.org/doi/10.5555/2227895.2227901.
2. Petrović S, Osborne M, McCreadie R, Macdonald C, Ounis I, Shrimpton L. can twitter replace newswire for breaking news? In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. Cambridge, MA, United States: Association for the Advancement of Artificial Intelligence; 2013;713-716. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/14450.
3. Mamo N, Azzopardi J, Layfield C. Who? What? Event tracking needs event understanding. In: Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - (Volume 1). Remote: SciTePress; 2021;139-146. Available from: https://www.scitepress.org/PublicationsDetail.aspx?ID=o+lys1RHmPU=&t=1.
4. Mamo N, Layfield C, Azzopardi J. from event tracking to event modelling: understanding as a paradigm shift. In: Discovery Knowledge, editor. Knowledge engineering and knowledge management. Remote: Springer Cham; 2023. p. 21–36.
5. Maldonado A, Lewis D. Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In: Proceedings of the 12th International Conference on Terminology and Knowledge Engineering. Copenhagen, Denmark: Copenhagen Business School; 2016;91-100. Available from: http://hdl.handle.net/2262/82537.

6. Simon N, Kešelj V. Automatic term extraction in technical domain using part-of-speech and common-word features. In: DocEng '18: Proceedings of the ACM Symposium on Document Engineering 2018. Halifax, Nova Scotia, Canada: Association for Computing Machinery; 2018;1–4. Available from: https://dl.acm.org/doi/abs/10.1145/3209280.3229100.
7. Lopes L, Fernandes P, Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency -TF-DCF. Knowl-Based Syst. 2016;97:237–49.
8. Zhang Z, Gao J, Ciravegna F. SemRe-rank: improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. ACM Trans Knowl Discov Data. 2018;12(5):1–41.
9. Loukachevitch N, Nokel M. An experimental study of term extraction for real information-retrieval thesauri. In: Proceedings of the 10th International Conference on Terminology and Artificial Intelligence. Paris, France; 2013;71–78.
10. Olteanu A, Castillo C, Diaz F, Vieweg S. CrisisLex: a lexicon for collecting and filtering microblogged communications in crises. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. Ann Arbor, MI, USA: Association for the Advancement of Artificial Intelligence; 2014;376-385. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/14538.
11. Temnikova I, Castillo C, Vieweg S. EMTerms 1.0: a terminological resource for crisis tweets. In: 12th Proceedings of the International Conference on Information Systems for Crisis Response and Management. Krystiansand, Norway: University of Agder (UiA); 2015. p. 134-146. Available from: http://idl.iscram.org/files/irinatemnikova/2015/1229_IrinaTemnikova_etal2015.pdf.
12. Hossny AH, Mitchell L. Event detection in Twitter: a keyword volume approach. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). Singapore: IEEE; 2018;1200-1208. Available from: https://ieeexplore.ieee.org/document/8637560.
13. Hua T, Chen F, Zhao L, Lu CT, Ramakrishnan N. Automatic targeted-domain spatiotemporal event detection in Twitter. GeoInformatica. 2016;20(4):765–95.
14. Zhou D, Chen L, Zhang X, He Y. Unsupervised event exploration from social text streams. Intell Data Anal. 2017;21(4):849–66.
15. Saeed Z, Abbasi RA, Razzak I, Maqbool O, Sadaf A, Xu G. Enhanced heartbeat graph for emerging event detection on twitter using time series networks. Expert Syst Appl. 2019;136:115–32.
16. Meladianos P, Nikolentzos G, Rousseau F, Stavrakas Y, Vazirgiannis M. Degeneracy-based real-time sub-event detection in twitter stream. In: Proceedings of the Ninth International AAAI Conference on Web and Social Media. Oxford, United Kingdom: The AAAI Press; 2015;248-257. Available from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewPaper/10502.
17. Hasan M, Orgun MA, Schwitter R. Real-time event detection from the twitter data stream using the TwitterNews+ Framework. Inf Process Manage. 2019;56(3):1146–65.
18. Saeed Z, Abbasi RA, Maqbool O, Sadaf A, Razzak I, Daud A, et al. What's happening around the world? A survey and framework on event detection techniques on Twitter. J Grid Comput. 2019;17(2):279–312.
19. McMinn AJ, Jose JM. Real-time entity-based event detection for Twitter. In: Mothe J, Savoy J, Kamps J, Pinel-Sa, Pinel-Sauvagnat K, Jones G, et al., editors. CLEF 2015: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Toulouse, France: Springer International Publishing; 2015;65–77. Available from: https://link.springer.com/chapter/10.1007/978-3-319-24027-5_6.
20. Chen X, Li Q. Event modeling and mining: a long journey toward explainable events. VLDB J. 2020;29(1):459–82.
21. Farzindar A, Khreich W. A survey of techniques for event detection in Twitter. Comput Intell. 2015;31(1):132–64.
22. Chung TM. A corpus comparison approach for terminology extraction. Terminology. 2003;9(2):221–46.
23. Makkonen J, Ahonen-Myka H, Salmenkivi M. Simple semantics in topic detection and tracking. Inf Retrieval. 2004;7(3):347–68.
24. Li B, Li W, Lu Q, Wu M. Profile-based event tracking. In: SIGIR '05: Proceedings of the 28th ACM/SIGIR International Symposium on Information Retrieval 2005. Salvador, Brazil: ACM; 2005;631–632. Available from: http://dl.acm.org/citation.cfm?id=1076163.
25. De Boom C, Van Canneyt S, Dhoedt B. Semantics-driven event clustering in Twitter feeds. In: Proceedings of the 5th Workshop on Making Sense of Microposts. Florence, Italy: CEUR; 2015;2-9. Available from: http://ceur-ws.org/Vol-1395/.
26. Chen HH, Ku LW. In: Allan J, editor. An NLP & IR approach to topic detection. vol. 12. Topic detection and. tracking. Boston: Springer; 2002. p. 243–364.
27. Luo Z, Wang H, Xie R. Extract domain terminologies for knowledge graph construction using domain feature vectors. In: 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA). Beijing, China: IEEE; 2017;53-57. Available from: https://ieeexplore.ieee.org/document/8078715.
28. Medelyan O, Witten IH, Divoli A, Broekstra J. Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. Wiley Interdiscip Rev Data Min Knowl Discov. 2013;3(4):257–79.
29. Mamo N, Azzopardi J, Layfield C. Fine-grained topic detection and tracking on Twitter. In: Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - (Volume 1). Remote: SciTePress; 2021;79-86. Available from: https://www.scitepress.org/PublicationsDetail.aspx?ID=o+Iys1RHmPU=&t=1.
30. Reed JW, Jiao Y, Potok TE, Klump BA, Elmore MT, Hurson AR. TF-ICF: a new term weighting scheme for clustering dynamic data streams. In: 2006 5th International Conference on Machine Learning and Applications (ICMLA'06). Orlando, Florida, USA: IEEE; 2006;258–263. Available from: https://ieeexplore.ieee.org/document/4041501.
31. Samant SS, Bhanu Murthy NL, Malapati A. Improving term weighting schemes for short text classification in vector space model. IEEE Access. 2019;7:166578–92.
32. Park Y, Patwardhan S, Visweswariah K, Gates SC. An empirical analysis of word error rate and keyword error rate. In: INTERSPEECH 2008: Proceedings of the 9th Annual Conference of the International Speech Communication Association. Brisbane, Australia; 2008;2070–2073. Available from: https://www.isca-speech.org/archive_v0/interspeech_2008/i08_2070.html.

33. Kit C, Liu X. Measuring mono-word termhood by rank difference via corpus comparison. Terminol Int J Theore Appl. 2008;14(2):204–29.
34. Bird S, Loper E, Klein E. Natural language processing with Python. 1st ed. O'Reilly Media Inc.; 2009. Available from: https://www.nltk.org/book/.
35. UEFA.: Dictionary-inside UEFA. Accessed on 11 Aug 2021. Available from: https://www.uefa.com/insideuefa/dictionary/index.html.
36. Wikipedia: Glossary of Association Football Terms. https://en.wikipedia.org/wiki/Glossary_of_association_football_terms. Accessed 11 Aug 2021.
37. Kubo M, Sasano R, Takamura H, Okumura M. Generating live sports updates from Twitter by finding good reporters. In: Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). vol. 1. Atlanta, Georgia, USA: IEEE Computer Society; 2013;527–534. Available from: http://dl.acm.org/citation.cfm?id=2568811.
38. Zhang J, Yao Jg, Wan X. Towards constructing sports news from live text commentary. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics; 2016;1361–1371. Available from: https://www.aclweb.org/anthology/P16-1129.
39. Formula 1.: F1 Glossary—A–Z List of the Top Formula 1 Terms. https://www.formula1.com/en/championship/inside-f1/glossary.html. Accessed 12 Aug 2021.
40. F1technical.: formula one glossary. https://www.f1technical.net/glossary/. Accessed 12 Aug 2021.
41. Formula 1 Dictionary.: Formula 1 Dictionary. formula1-dictionary.net. Accessed 12 Aug 2021.
42. Wikipedia.: glossary of motorsport terms. https://en.wikipedia.org/wiki/Glossary_of_motorsport_terms. Accessed 12 Aug 2021.
43. Dole Institute of Politics.: Political Glossary. https://doleinstitute.org/get-involved/civic-engagement-tools/political-glossary/. Accessed 14 Jul 2021.
44. Brown GW, McLean I, McMillan A. The Concise Oxford Dictionary of Politics and International Relations. 4th ed. Oxford Quick Reference. Oxford: Oxford University Press; 2018. Available from: https://www.oxfordreference.com/view/10.1093/acref/9780199670840.001.0001/acref-9780199670840.
45. Justia.: criminal law glossary. https://www.justia.com/criminal/glossary/. Accessed 14 Jul 2021.
46. United States Courts.: glossary of legal terms. https://www.uscourts.gov/glossary. Accessed 14 Jul 2021.
47. U S Election Assistance Commission.: Glossary of Terms Database. https://www.eac.gov/glossary. Accessed 14 Jul 2021.
48. English First.: 1000 Most Common Words in English. https://www.ef.com/wwen/english-resources/english-vocabulary/top-1000-words/. Accessed 16 Jul 2021.

Publisher's Note