Expert Systems With Applications 118 (2019) 522–538

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A framework for event classification in tweets based on hybrid semantic enrichment

Simone Romero*, Karin Becker

Institute of Informatics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 17 January 2018
Revised 27 September 2018
Accepted 14 October 2018
Available online 15 October 2018

Keywords:
Event classification
Semantic web
DBPedia
Twitter
Discriminative features

ABSTRACT

Twitter has become instrumental as a means of spreading information, opinions or awareness about real-world events. The classification of event-related tweets is a challenging problem since tweets are noisy and sparse pieces of text that lack contextual information. Related work proposes contextual enrichment techniques using external sources (e.g. semantic web, external documents), often considering underlying assumptions about the target events. However, they lack guidelines for determining the textual features to enrich, the external sources to use, the properties to explore, and how to prevent the inclusion of unrelated information. In this paper, we propose a hybrid semantic enrichment framework for the classification of event-related tweets. We contribute to this field by leveraging different contextual enrichment strategies into a unifying framework targeted at a broad range of event types, where each enrichment technique has a role in the improvement of event classification. The framework also encompasses a solution to deal with the huge number of features that result from semantic enrichment, which combines a pruning method to select domain relevant semantic features and general-purpose feature selection techniques. We assessed the contribution of each framework component to event classification improvement using a broad experimental setting. Using seven events of distinct natures, we outperformed a word embeddings baseline in 93.6% of cases, and a textual baseline in 60.3% of cases. In most cases, we improved the recall, with no significant impact on the precision.

1. Introduction

Social media platforms have become essential as a means of spreading information, opinions, or awareness about real-world events. Twitter is a popular platform, where a huge number of messages about all sorts of topics and subjects is published every day, thus becoming a primary information source about events of all types and magnitudes (Li, Nourbakhsh, Shah, & Liu, 2017; Sakaki, Okazaki, & Matsuo, 2010). The identification and classification of event-related tweets have many useful applications, such as the identification of breaking news, rumors, and fake news; real-time awareness and updating about events (e.g. car crashes, political protests, fires, epidemics, sports, culture, urban daily events); crisis management; measurement of the consequences of a given event (e.g. volume of messages, sentiment); amongst others (Khare, Burel, & Alani, 2018; Liu et al., 2016; Saif, He, & Alani, 2012; Schulz, Ristoski, & Paulheim, 2013; Singh, Dwivedi, Rana, Kumar, & Kapoor, 2017).

However, the classification of event-related tweets is a complex problem that cannot be handled solely by conventional Natural Language Processing (NLP) techniques. By design, Twitter messages contain an informal and dynamic vocabulary, and several tricks are used to convey meaning (e.g. hashtags, URLs, emoticons). Thus, tweets are very noisy text pieces, in which the lack of context can make the task of classifying them very difficult. In addition, Twitter users post messages massive scale with a variety of content types, which differ in subject, scope, and purpose (Chen, Zhou, Sellis, & Li, 2018; Schulz et al., 2013). As a result, it is not easy to classify event-related tweets within such a huge volume of heterogeneous and sparse data. Thus, this problem requires advanced techniques from fields such as artificial intelligence, information retrieval, and data mining.

The classification of events in tweets has drawn increasing attention in the expert and intelligent systems field. Regardless of the event scale, government, individuals, as well as private and public organizations, can take advantage of the correct classification of event-related tweets to support all levels of decision making. The awareness and monitoring of events can be leveraged to support decisions to enhance public safety, reduce the impact of negative events on the population, obtain economic gains,

* Corresponding author.
E-mail addresses: sapromero@inf.ufrgs.br (S. Romero), karin.becker@inf.ufrgs.br (K. Becker).

https://doi.org/10.1016/j.eswa.2018.10.028
0957-4174/

promote culture and sports, prevent misinformation, detect fake news, contribute to smart cities, among others. Examples of end-to-end intelligent systems that incorporate event-related tweet classification are Reuters Tracer (Liu et al., 2017) and TwitterStand (Sankaranarayanan, Samet, Teitler, Lieberman, & Sperling, 2009) for detecting breaking news; Twitcident for monitoring fire events (Abel, Hauff, Houben, Stronkman, & Tao, 2012); and ArmaTweet, targeted at crisis management of terrorism events (Tonon, Cudré-Mauroux, Blarer, Lenders, & Motik, 2017).

The Event Classification problem deals with the construction of classification models to filter and categorize event-related tweets using learning techniques on data. It can target specific events (e.g. Rio Olympics), or events within a category/domain (e.g. sportive events) (Adedoyin-Olowe, Gaber, Dancausa, Stahl, & Gomes, 2016; Atefeh & Khreich, 2015). Event classification is often a complex and important part of a broader system that identifies events in never-ending and evolving data streams (Chen et al., 2018; Li et al., 2017; Liu et al., 2016; Packer, Samangooei, Hare, Gibbins, & Lewis, 2012; Sakaki et al., 2010).

Existing approaches for event classification in tweets often focus on specific types of events, such as epidemics (Fisichella, Stewart, Cuzzocrea, & Denecke, 2011), incidents (e.g. car crashes, fires) (Abel et al., 2012), or natural disasters (Sakaki et al., 2010). They adopt particular definitions of events that rely on assumptions about the tweets, such as volume of posts (Singh et al., 2017), temporal and geospatial properties (e.g. small-scale incidents vs. massive crisis situations) (Anantharam, Barnaghi, Thirunarayan, & Sheth, 2015; Schulz et al., 2013), retweeting behavior (Chen et al., 2018), properties of the vocabulary used (Adedoyin-Olowe et al., 2016; Becker, Iter, Naaman, & Gravano, 2012; Medvet & Bartoli, 2012), or existence of prior information about the event (Packer et al., 2012). A major difficulty in all of these approaches is to transpose the proposed techniques to a new event classification context, in which the original assumptions may not hold. As a side effect, the performance of distinct event classification approaches cannot be compared.

A common feature in all of these works is that they rely on contextual enrichment techniques as a means of dealing with the poor textual contents of tweets. Proposed techniques for adding contextual information include interpolation of related external documents (Chen et al., 2018; Genc, Sakamoto, & Nickerson, 2011; Rosa, Shah, Lin, Gershman, & Frederking, 2011; Sicilia, Giudice, Pei, Pechenizkiy, & Soda, 2018; Vosecky, Jiang, Leung, Xing, & Ng, 2014), knowledge from the semantic web to generalize their contents (Abel et al., 2012; Khare et al., 2018; Packer et al., 2012; Rowe & Stankovic, 2011; Schulz, Guckelsberger, & Janssen, 2015; Schulz et al., 2013; Tonon et al., 2017), extraction of Named Entities related to categories of interest (Abel et al., 2012; Rowe & Stankovic, 2011; Saif et al., 2012; Vosecky et al., 2014), and distributional semantics extracted from large corpora (Edouard, Cabrio, Tonelli, & Le Thanh, 2017; Li, Shah, Liu, Nourbakhsh, & Fang, 2016; Wang & Niu, 2018). However, these techniques have not been evaluated with regard to their contribution to the event classification problem, because they are often part of a solution targeted at events with specific properties (e.g. natural disasters, small-scale incidents, bursty events). There is a lack of guidelines for determining the textual features to enrich, the external sources to use, the properties to explore, and how to prevent the inclusion of unrelated information. In addition, contextual enrichment results in a huge amount of new features, most of which have no discrimination power for the event classification task (Janpuangtong & Shell, 2015; Khare et al., 2018; Ristoski & Paulheim, 2016; Romero & Becker, 2016; Schulz et al., 2015).

In this paper, we propose a framework for the classification of event-related tweets. The distinctive features of this framework are: (a) it is agnostic concerning to the properties of events; (b) it explores a hybrid contextual enrichment strategy; and (c) it addresses the selection of discriminative features for event classification. The framework assumes a definition that encompasses a broad range of events and event types. For our purposes, *an event is an occurrence, represented by a topic, related to a specific time, that can involve one or more locations and agents* (Romero, 2017).

We propose a hybrid contextual enrichment process, which combines semantic enrichment, external documents, and Named Entity Recognition (NER). These three techniques are complementary in their role providing context: (a) NER aims to recognize, in tweets and external documents, the entities that are relevant for event characterization; (b) external document enrichment helps overcome the poor and sparse textual contents of tweets with more structured and rich related textual properties; and (c) semantic enrichment leverages the semantic properties available in the Linked Open Data (LOD) cloud (Bizer, Heath, & Berners-Lee, 2009; Schmachtenberg, Bizer, & Paulheim, 2014), so as to obtain domain representative concepts.

Since semantic enrichment results in a vast number of features (Janpuangtong & Shell, 2015; Ristoski & Paulheim, 2016; Schulz et al., 2015), our framework combines a pruning method for the selection of relevant semantic features with general purpose feature selection methods (Hall, 1999). We developed a pruning algorithm based on PageRank (Page, Brin, Motwani, & Winograd, 1999) to discard the concepts that are either too generic or too specific and thus not representative of the event domain. General feature selection methods (e.g. based on information gain or statistical properties) address the complementary problem of selecting discriminative features with regard to the classification task.

Our experiments aim to evaluate the framework performance for event classification and the contribution of each of its underlying components. For this reason, the experimental setting includes: (a) seven datasets representing events of distinct natures (e.g. sports, natural disasters, epidemics, celebration date); (b) a performance comparison involving the contextual enrichment techniques; (c) a performance comparison of the techniques to identify relevant features for event classification; (d) three different classification algorithms widely used in text classification problems, and (e) two different baselines. With regard to our early work (Romero & Becker, 2016; 2017), we evolved the underlying components of the framework and developed an in-depth analysis of the contribution of its components to the improvement of event-related tweet classification.

Our contributions can be summarized as follows:

- A framework for the classification of events in tweets that encompasses a broad range of events and event types, such that one needs not to be concerned with underlying assumptions such as volume of posts (Singh et al., 2017), temporal and geospatial properties (Schulz et al., 2015), social behavior (Chen et al., 2018), existence of prior information about the event (Packer et al., 2012), etc.;
- A hybrid semantic enrichment process that extracts textual features from both tweets and referenced documents, and aggregates semantics by identifying named entities and knowledge from the LOD cloud. We leverage enrichment techniques proposed in distinct related work into a unifying framework, and evaluate their contribution to the event classification problem;
- The combination of a pruning method to select domain-related semantic features, and general-purpose feature selection methods to improve the performance of the classification task. Related work has recognized the huge number of irrelevant features resulting from semantic enrichment, but has addressed this issue with feature selection techniques of which the value has not been evaluated (Khare et al., 2018; Romero & Becker, 2016; Schulz et al., 2015);

- The evaluation of each underlying component of the proposed framework considering a broad experimental setting that includes different events, data preparations and classification algorithms;
- The comparison of the proposed framework with an alternative enrichment approach based on distributional semantics (i.e. word embeddings), a state-of-the-art approach for improving text classification results in general.

Our experimental results show that the hybrid semantic enrichment significantly improves the classification of event-related tweets. We outperformed the word embeddings baseline in 93.6% of cases, showing that the proposed framework based on hybrid semantic enrichment is comparable to alternative state-of-the-art techniques. With regard to the textual baseline, we achieved better results in 60.3% of cases, mainly in terms of Recall. We conclude that semantic enrichment adds concepts representative of the domain, and external documents enrichment helps to overcome the poor and sparse textual contents of tweets with related textual properties. Our experiments also confirmed the complementary roles of the proposed pruning algorithm and general-purpose feature selection techniques. The former is essential to benefit from semantic enrichment, as it contributes to a better balance between textual and semantic features (50% each). The latter contributes to the classification task by selecting discriminative features with regard to the classification (Romero, 2017). The result of this combination is the improvement in the recall, with no or minor impact on precision. The recall is an important metric in this context as indicates the ability to recognize the event-related tweets as such.

The rest of this paper is structured as follows: Section 2 presents an overview of related work. Section 3 highlights the main aspects of the proposed approach. Section 4 provides a detailed description of the experiments and their results. Conclusion and future work are addressed in Section 5.

2. Related work

Event Identification and Classification aims at identifying tweets related to events that are either specific, or within a category/domain (Adedoyin-Olowe et al., 2016; Atefeh & Khreich, 2015; Sakaki et al., 2010). *Event identification* addresses the overall task of creating groups of subject and time-related tweets from large, never-ending data streams, using techniques such as topic modeling and social behavior analysis (Chen et al., 2018; Li et al., 2017; Liu et al., 2016; Packer et al., 2012; Sakaki et al., 2010). *Event classification* is a specific task within this context, that deals with the construction of classification models to filter and categorize events (Abel et al., 2012; Khare et al., 2018; Schulz et al., 2015; Schulz et al., 2013). In this paper, we assume such a distinction and focus on the *event classification* problem.

One of the key challenges in event-related tweet classification is the extraction of relevant features from such a sparse collection of short and noisy pieces of text. Related work employs contextual enrichment to convey useful meaning for the classification task using four basic approaches: (a) *named entity recognition*; (b) *external documents*; (c) *semantic enrichment*; and (d) *distributional semantics. Hybrid* approaches combine different strategies.

The most straightforward approach is to employ NER tools (e.g. Open Calais, Alchemy) to identify named entities that are mentioned in documents (Abel et al., 2012; Rowe & Stankovic, 2011). This technique contextualizes textual features according to categories of interest (e.g. person, organization, location). This approach is very popular and it is often combined with the other approaches described below.

External document enrichment combines the textual features from the tweets and related external documents. Works in this category vary in the external source of information (e.g. referenced URLs, Wiki pages, other tweets), and the features extracted from these sources. Chen et al. (2018) characterize events according to the retweeting behavior and combine textual content, hashtags, time, and location (in tweets and retweets) in high volumes of related and evolving tweets. Sicilia et al. (2018) present a rumor detection system for Twitter focused on the health domain. They extract statistical and sentiment features from tweets, referred URLs, and retweets, in addition to conversation size and the likelihood of retweets. These works assume specific properties about events, such as domain or volume of tweets. Vosecky et al. (2014) extract specific features from tweets (frequent co-occurrences of hashtags and terms) and external documents (named entities and top-k frequent terms). Rosa et al. (2011) extract all textual features from URLs referred within tweets, but the excess of features degraded the performance of the event classifiers. Genc et al. (2011) employ Wikipedia articles for topic categorization, by checking for each word contained in a tweet whether a specific Wiki page exists. Extracting all textual features from external documents can be very time-consuming, and do not necessarily result in discriminative features. Therefore, these propositions need to be complemented with strategies for the selection of the relevant terms to be extracted and techniques to assess their contribution to the event classification task.

Semantic enrichment leverages knowledge from the Semantic Web, such as LOD cloud datasets (e.g. DBPedia, Geonames, YAGO) or social knowledge bases (e.g. socialbakers). These knowledge sources describe resources using properties that belong to specific vocabularies (e.g. rdfs, skos). Features extracted from tweets are mapped into these resources and their properties retrieved, resulting in semantic features that generalize and complement tweet contents. Each work in this category varies in the knowledge source and properties explored. ArmaTweet (Tonon et al., 2017) focuses on security-related events. It maps tweet terms into DBPedia and Wordnet resources to create a full-fledged semantic description of events of interest, which can be used to query tweets using SPARQL. The works developed by Schulz et al. (2015, 2013) focus on small-scale incidents (e.g. car crashes, fires), which do not rely on bursts of vocabulary. They extract from tweets spatial, temporal, and TF-IDF weighted terms (Schulz et al., 2013). The authors also experimented with named entities and advanced heuristics for identification of temporal and spatial expressions (Schulz et al., 2015). These features are then related to resources of DBPedia, using different properties, such as *rdfs:subClassOf, skos:broader*, and *dct:subject*. Their best results are obtained using spatial heuristics, followed by features extracted using NER tools. Rowe and Stankovic (2011) use Zemanta to map tweet named entities into DBPedia concepts using the *dct:subject* property. Khare et al. (2018) combine statistical and semantic features for classifying tweets across different types of crisis events. Semantic features are extracted from BableNet and DBPedia using different properties (e.g *rdf:type, rdfs:label, dct:subject, dbo:country*). All of these works target different types of events and do not provide guidelines (nor justifications) for the selection of the chosen knowledge bases or properties explored. The assessment of the contribution of the semantic enrichment is thus hidden in a broader classification solution targeted at events with specific properties. Although these works recognize that classification performance can be degraded by the huge number of resulting semantic features, this issue is addressed only by Khare et al. (2018), who use a feature selection technique.

In an earlier version of our work (Romero & Becker, 2016), we compared the semantic enrichment of named entities (agents, location) and terms (frequent and representative terms). We used DBPedia and the *rdf:type* property. We observed that DBPedia showed good coverage, but semantic enrichment yielded very

sparse and high dimensional datasets. We experimented two feature selection algorithms, but the results were very dependent on the classification algorithm. Janpuangtong and Shell (2015) addressed the problem of selecting relevant domain concepts from the LOD cloud using a centrality-based approach based on HITS. This technique was originally proposed to find relevant public datasets for data mining, but it could be transposed to filter out relevant features resulting from semantic enrichment.

Hybrid enrichment combines the aforementioned techniques, where each one plays a different role in providing contextual information. NER is the primary technique used in most works described above. Others combine external documents with semantic enrichment. Packer et al. (2012) assumes the existence of prior information about specific events (e.g.web pages), which are used to extract terms to be enriched using YAGO2 knowledge base (*rdfs:label*) to filter relevant tweets. Abel et al. (2012) propose a framework for filtering, searching, and analyzing information about real-world incidents (e.g. fire, car crashes). Their framework is connected to an emergency broadcasting service, from which named entities are extracted and enriched using DBPedia. The evaluation of the framework shows that the semantic enrichment boosts the filtering performance of tweets related to a specific type of incident. However, all of these systems assume that prior information on the event or event type is available.

More recently, distributional semantics extracted from large corpora has been explored. It consists of extracting multidimensional representations of words that capture their similarity according to the context (i.e. embeddings), which can be treated as an alternative form of contextual enrichment. In the context of Reuters Tracer system for detecting breaking news (Liu et al., 2017), Li et al. (2016) combined semantic enrichment using a light-weight knowledge base and word embeddings to classify tweets in real time, resulting in superior performance when compared to semantic enrichment only. Edouard et al. (2017) exploit information acquired from the LOD cloud (YAGO, DBPedia) to enrich named entities mentioned in event-related tweets. This enriched content is then used to extract features and build word-embedding vectors, which are used to train models for event detection and classification. Wang and Niu (2018) propose a climate event detection algorithm based on domain-specific word embedding. They built the climate document representation model using continuous-bag-of-words, combined with other factors, such as climate category, occurrence time, and occurrence position. Although these works presented good results, they rely on the existence of a huge corpus, preferably domain-specific. Training domain specific word embeddings model over tweet texts can be very time-consuming.

Table A.11 in Appendix A summarize how these works vary on the features selected, enrichment technique employed, choice of relevant features, and learning algorithm applied.

We contribute to this field by leveraging different contextual enrichment strategies into a unifying framework targeted at a broad range of event types, where each enriching technique contributes to the improvement of event classification in a different way. The framework also addresses the problem of selecting semantic features that are relevant and discriminative for the classification task. Initial ideas were presented in our earlier work (Romero & Becker, 2017), which are further detailed and assessed in this paper.

3. A hybrid enrichment framework for event classification in tweets

Fig. 1 presents the proposed framework for event classification in tweets. The striking features are:

- *Core features*: guided by the event definition adopted, we extract from documents the following core features (Romero & Becker, 2016): (a) *Vocabulary*, i.e. terms that represent the topic or subject of the event, either because they are frequent or representative; (b) *Agents* (e.g. people, organization) that are involved in the event, or affected by it; and (c) *Location* that represents geographic information about the event. These features correspond to the *what, who* and *where* of events (Liu et al., 2016);
- *Hybrid semantic enrichment*: knowledge extracted from the LOD cloud is the central approach for adding concepts that characterize the event domain. To boost its results, it is combined with external document enrichment and NER. External related web documents referenced in tweets enable to overcome the poor and sparse textual contents, expanding the set of conceptual features. NER recovers from documents conceptual features from specific categories, namely Agents and Location.
- *Selection of relevant semantic features*: good results in the classification process are dependent on relevant textual and semantic features in the dataset. We propose to combine two methods: a semantic pruning mechanism based on PageRank, and a general-purpose feature selection technique. The former aims at discarding semantic features that are either too generic, too specific, and thus not representative of the domain. It is complementary to feature selection methods, which aim at finding features discriminative with regard to the classification task.

The framework provides the following benefits: (a) it guides the extraction of textual features that characterize a broad class of events; (b) it leverages existing enrichment techniques into a single unifying framework, each one with a specific contribution in the provision of contextual information; and (c) it deals with the dimensionality of the resulting dataset by deploying techniques that address the relevance of semantic features concerning to the domain and its discrimination power in the classification task.

The underlying process is divided into six steps, depicted in darker grey. First, all tweets are *pre-processed*, where traditional actions are taken. External related web documents are also recovered in this step, through the recognition of the URLs mentioned in the tweets. Then, we *extract conceptual features* from the content of the tweets and the web documents, to be semantically enriched in the next phase. Next, these conceptual features are *semantically enriched* using knowledge from the LOD cloud. The resulting semantic features are then *pruned*, in order to discard concepts that are either too generic or too specific. The pruned semantic features are then *incorporated* to the textual features extracted from the original tweets. Finally, these datasets are submitted as input to the *classification* step, preceded by a *feature selection* method. Fig. 2 depicts an example of the features extracted, enriched and pruned according to the proposed approach based on one of the datasets used in our experiments,[1] which will be used as running example.

3.1. Pre-processing

Given a set of tweets, the goal of this step is to perform basic pre-processing actions, such as tokenization, removal of re-tweets, and the normalization of specific features (i.e. @User, URLs, and emoticons to T_USER, T_URL, and T_EMOT). These data preparation techniques reduce the number of features in the dataset and address over-fitting issues. We also identify the URLs mentioned in the tweets and extract their contents.

[1] Influenza dataset.

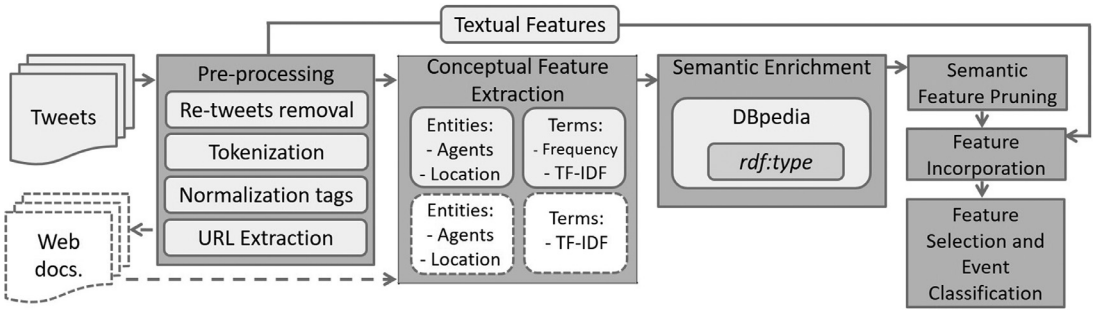

Fig. 1. Summarized pipeline of the Event Classification process.

Original tweet	@user: Officials worry about swine flu preparedness amid budget cuts. http://bit.ly/J2gZX	
Web document content	http://bit.ly/J2gZX: "... six years of worrying about bird flu did much to prepare the United States for the current swine flu outbreak."	
Processed tweet	T_USER: Officials worry about swine flu preparedness amid budget cuts T_URL	
Conceptual Features Extracted	**From tweet text**	**From URL**
	Swine (frequent/representative terms) Worry (representative term)	Flu (representative term) United States (location)

Semantic Features (rdf: type)	**Semantic Feature Pruning**	
	Too Generic	**Too Specific**
http://dbpedia.org/page/Swine_influenza owl:Thing, dbo:Disease, wikidata:Q12136 http://dbpedia.org/page/Influenza owl:Thing, dbo:Disease, umbel-rc:AilmentCondition http://dbpedia.org/page/United_States owl:Thing, dbo:PopulatedPlace yago:WikicatEnglish-speakingCountriesAndTerritories	owl:Thing	wikidata:Q12136 umbel-rc:AilmentCondition yago:WikicatEnglish-speakingCountriesAndTerritories
Pruned Semantic Features	dbo:Disease, dbo:PopulatedPlace	
Incorporated textual and semantic features		
T_USER, officials, worry, about, swine, flu, preparedness, amid, budget, cuts, T_URL, dbo:Disease, dbo:PopulatedPlace, ...		

Fig. 2. A running example according to the proposed framework.

The outputs of this step are: (a) the tweet uni-grams and normalized tokens (e.g. T_USER, T_EMOT), and (b) a list of the URLs extracted from the tweet text and their respective contents.

3.2. Conceptual feature extraction

The goal of this step is to extract from these pre-processed documents a set of conceptual features, which are the ones to be semantically enriched in the next step of the process. However, to ensure good results in the event classification, it is important to define which kind of features to extract, the source documents, as well as the corresponding extraction methods. Our framework addresses these issues by defining a set of core features types to be extracted, adopting external document enrichment to complement tweet texts, and identifying the extraction technique appropriate to each case.

3.2.1. Core feature types

As mentioned, the core features proposed correspond to the *what*, *who* and *where* of events (Fisichella et al., 2011; Liu et al., 2016). In our previous work, we concluded that the importance of these features is not related to the type of event (Romero & Becker, 2016).

- *Vocabulary*: Refers to terms that are frequent, representative of a domain, or created in bursts for a specific event (Anantharam et al., 2015; Becker, Naaman, & Gravano, 2011; Medvet & Bartoli, 2012; Sakaki et al., 2010; Schulz et al., 2015). This set of terms can also describe the topic or subject that the event refers to. The used vocabulary represents the popularity or impact caused by the event, and the scale defines representativeness of the tweets in which it can be detected;
- *Agent*: Broadly defined as people, organizations, products or services, can be identified in events of all natures, either in an active role (e.g. a brand subject to a marketing action, an artist in a cultural event, a policeman in an accident), or a passive one (e.g. people affected by a natural disaster or epidemic). While active agents can be previously defined in planned events, passive agents must be part of the event detection task otherwise (Sakaki et al., 2010);
- *Location*: Geographic property related to the event itself, to a user affected, or to a place from which it is reported (Anantharam et al., 2015; Becker et al., 2011; Schulz et al., 2015). Considering event detection approaches based on users as sensors, the location of the report or the user is quite relevant, and it is a synonym of the location of the event (Sakaki et al., 2010). However, only about 1% of all tweets contain geographic metadata, so quite often it is impossible to determine the origin of the post (Schulz et al., 2015). Therefore, location information needs to be identified in the text itself or from the user profile.

Scale and the temporal component (i.e. the *when*) are very important, but they are more related to the task of event identification. Thus, they are not considered in this work.

3.2.2. External documents

As mentioned, the adoption of external documents complement the contents of the tweets, in the search for conceptual features that are representative of event description. These related web documents are recognized through the URLs mentioned in the post and can represent another tweet or other type of web page (e.g. sportive news, blog).

3.2.3. Extraction techniques

We propose to identify Agents and Location using NER tools, since this kind of tool is the more indicated to recognize specific categories of terms presented in the text (Abel et al., 2012; Saif et al., 2012; Schulz et al., 2013). For the Vocabulary, we propose to adopt both *frequent* and *representative* terms. The former can be identified using the top-k frequent terms, in order to avoid sparsity, and the latter, using term weighting techniques, such as TF-IDF given a threshold (Schulz et al., 2013).

We do not extract the top-k frequent terms from external web documents, because the resulting list presents many terms not related to the domain of the event analyzed. In addition, experiments demonstrated that frequent and representative terms were very similar.

The example of Fig. 2 shows the entities/vocabulary extracted from the tweet text and referenced web page. Note that the vocabulary extraction strategy does not select a feature if it contributes to sparsity (e.g. "budget"), and that external documents allow extracting vocabulary relevant to the domain, but which might be sparse within the tweets set (e.g. "flu"). Therefore, combining features extracted from tweet texts and external web documents helped us to better represent the domain of the target event.

3.3. Semantic enrichment

Given a set of conceptual features as input, this step aims to generalize them to obtain more domain-representative concepts, by retrieving associated knowledge from the LOD cloud. The output is a set of semantic features.

This step involves two tasks: (a) *mapping* the feature into a resource in a specific knowledge base in the LOD cloud, and (b) *retrieving properties* of the resource, as described in the knowledge base. Different properties can be selected to assign meaning to tweets, such as *Type, Category* or *sameAs*.

Likewise, according to the application purpose, different knowledge bases can be deployed, such as DBPedia, Geonames, YAGO, etc (Packer et al., 2012; Schulz et al., 2015). In this paper, we opted for using DBPedia since it is a cross-domain knowledge base, which has connections to several other datasets. It covers a huge amount of information, allowing us to obtain knowledge from tweet contents independent of the event type.

Regarding the property to be extracted, we decided to use the *rdf:type*, which contains general information about the resource (Rowe & Stankovic, 2011; Schulz et al., 2015). The *rdf:type* property objects (also referred to as concepts or semantic features) are used as complementary information, to help the generalization of the knowledge about the conceptual feature.

The matching between the conceptual features and the resource in the knowledge base can be performed automatically (e.g. RapidMiner *DBPedia Spotlight operator*), or by an *ad hoc* method (e.g. SPARQL queries). Fig. 2 illustrates the semantic enrichment for our running example.

3.4. Semantic feature pruning

This step aims to reduce the volume of semantic features resulting from the previous one, by selecting the ones that add the proper level of generalization to the conceptual features extracted from documents (Janpuangtong & Shell, 2015). This pruning phase is applied to the semantic features only, and the result is a reduced set of semantic features to be incorporated into the textual features from the dataset.

We implemented a variation of the PageRank method to analyze the relevance of the interlinked concepts extracted from the knowledge base, and discard the ones that are either too generic, too specific, or irrelevant to the domain. This type of feature tends to degrade the classification performance because they either introduce sparseness in the dataset or can be used to describe any situation besides the specific events to be characterized.

Considering the running example of Fig. 2, we can observe semantic features that are very representative of epidemic events, such as *Disease*. However, others are either too specific (e.g. *AilmentCondition*), or generic (e.g. *Thing*), and are discarded by our algorithm.

3.4.1. Algorithm description

PageRank was originally developed for rating web documents based on the link structure of the Web (Page et al., 1999). Using forward links (*inedges*) and backlinks (*outedges*), a random surfer visits these pages computing their salience in the graph. Pages with high scores are regarded as representative, where the importance of a page is defined in terms of inedges from other high score pages.

We adopt this idea by considering each node as a concept extracted from the LOD cloud, related by super/subclass relationships. The more general the concept, the highest the score. Likewise, the lowest the score, the more specific is the concept. Thus, the idea is to calculate the salience scores and to prune the ones with scores that are too high/low. Fig. 3 summarizes the proposed method: (a) the graph is initially constructed using the results of the previous step. The initial nodes are the LOD resources into which the vocabulary/entities were mapped; (b) more nodes are connected using the forward links, according to the other LOD concepts retrieved using the chosen properties (i.e. *rdf:type*); (c) these nodes are interconnected using super/subclass relationships also retrieved from the LOD cloud; and (d) given the built graph, PageRank is used to calculate the salience scores. Then, specific thresholds are applied to remove the nodes with highest and lowest scores, in order to select the semantic features that are potentially discriminative of the event to be characterized (depicted in black).

The pseudo-code is listed in Algorithm 1. Lines 3–6 create a graph using the resources, their respective concepts, the sub/superclass relationships between these concepts (i.e. *rdfs:SubClassOf*), as well as the number of sub/superclasses. In line 7, the scores are calculated following the traditional PageRank technique to build the graph. Line 8 corresponds to the actual pruning, where the features below/above the thresholds are discarded.

3.4.2. Automatic threshold definition

The graph produced by the pruning algorithm presents classes with a huge amount of relationships, as well as classes without any relationship. Thus, defining a threshold that meets all the event types and the characteristics of their semantic features is a challenging problem. Manually defining the threshold is an almost impossible task, since it requires in-depth knowledge about the event analyzed and the resources used for semantic enrichment. A more realistic possibility is to employ the score distribution of the concepts recognized in the knowledge base in combination with statistics measures, such as the median, quartiles, and interquartile range (IQR). The rationale behind it is that by using the median and the equations based on it, we can deal with these distorted

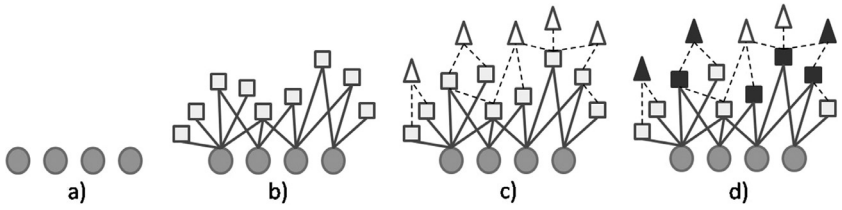

Fig. 3. Pruning concepts.

Algorithm 1 Page Rank-based feature pruning.

Input: $Concepts$ = concepts extracted from the LOD cloud
Output: $\emptyset$= set of relevant concepts
$Property \leftarrow$ **getTypes**($Concepts$)
$SubClass \leftarrow$ **getSubClassOf**($Property$)
$CountSC \leftarrow$ **getCountSubClassOf**($SubClass$)
$Matrix \leftarrow$ **getAdjacencyMatrix**($SubClass$)
$PRgraph \leftarrow$ **calculatePageRank**($Matrix$)
$\emptyset \leftarrow$ **performPruning**($CountSC, PRgraph$)
return $\emptyset$

values since the median is a measure that is not influenced by very large or minimal values.

We experimented with different strategies to automatically defining the pruning upper/lower thresholds. The best results, reported here, were obtained by the strategy referred to as Quartiles, which considered the values of the upper/lower quartiles of both the PageRank salience scores distribution, and the number of sub/superclass distribution. Other strategies are reported and evaluated in Romero (2017).

The QUARTILES strategy (Algorithm 2) was devised because by using the interconnection among the semantic features to calculate the salience score (i.e. PRgraph), we are considering only the network composed of the concepts retrieved according to the domain of the event analyzed. Thus, we obtain the relevance of these semantic features considering a limited portion of the LOD cloud. On the other hand, when we employ the number of sub/superclass of each concept to obtain the *node scores* (i.e. CountSC), we are also considering the influence of this concept in the whole LOD cloud. Hence, we can better discriminate among the relevant semantic features.

3.5. Feature incorporation

In this step, we aim at incorporating the pruned semantic features and the textual tokens from the tweet, to produce the training dataset for the Classification step. We adopted the Augmentation Method proposed by Saif et al. (2012) because we are able to maintain the original conceptual feature and additionally to include generalized information about it.

The training dataset contains both the textual tokens and the pruned semantic features from the LOD cloud. As presented in the running example in Fig. 2, the textual tokens extracted from the tweet text are incorporated with the pruned semantic features resulting from the previous step.

3.6. Feature selection and event classification

The goal of this step is to train the resulting incorporated dataset using classification algorithms. However, even after the execution of Semantic Feature Pruning step, the datasets can still contain a lot of semantic and textual features, to which the classification algorithms can be more or less sensitive. Thus, this step also assumes that other feature selection techniques can be applied as an attempt to further reduce the number of features and achieve better results in the Classification step.

We selected the *CfsSubsetEval* algorithm (Hall, 1999), which considers not only the relation with the target class, but also the redundancy among the features themselves. We also experimented with the InfoGain algorithm, which only the relationship with the target class, but the results were much inferior (Romero, 2017).

Finally, the prepared dataset is used as input to a supervised machine learning algorithm, in order to distinguish between the positive (i.e. target event) and negative examples (i.e. a non-event or an event different from the target). The event classification model is the final output of the process.

Algorithm 2 The QUARTILES strategy for pruning threshold definition.

Input: $PRgraph$ = graph of the concepts and its scores, $CountSC$ = number of sub/superclass of each concept
Output: $\emptyset$= set of relevant concepts
$QuartilesPR \leftarrow$ **calculateQuartile**($PRgraph$)
$QuartilesSC \leftarrow$ **calculateQuartile**($CountSC$)
for i in $PRgraph$ **do**
  $PRvalue \leftarrow PRgraph[i].value$
  $SCvalue \leftarrow CountSC[i].value$
  if ($PRvalue > QuartilesPR.Qi$ AND $PRvalue < QuartilesPR.Qs$) **then**
    if ($SCvalue > QuartilesSC.Qi$ AND $SCvalue < QuartilesSC.Qs$) **then**
      $\emptyset \leftarrow PRgraph[i]$
return $\emptyset$

Table 1
Description of the target datasets.

Dataset	# Tweets	Description	Event type
FaCup	1502	Final of the football season in England in 2012 [a].	Sports event
Olympics	1036	Women's gymnastics at Olympic Games (Sintsova, Musat, & Pu, 2013).	Sports event
Halloween	1551	Tweets about halloween.	Commemorative date
HSandy	1516	Hurricane Sandy that hit New York in 2012 [b].	Natural disaster
Alberta Flood	950	Tweets posted during 2013 Alberta floods [c].	Natural disaster
Australia Bushfire	881	Tweets posted during 2013 Australia bushfire [c].	Natural disaster
Influenza	1380	Tweets about Influenza (Lamb, Paul, & Dredze, 2013).	Epidemic

[a] http://www.socialsensor.eu/results/datasets/72-twitter-tdt-dataset.
[b] https://github.com/pavan046/benchmark-events-tweets-dataset.
[c] http://crisislex.org/tweet-collections.html.

4. Evaluation experiments

In this section, we describe the two main experiments performed to evaluate the contribution of the proposed framework. The first one (*Experiment #1*) aims at analyzing the performance of each component of the hybrid semantic enrichment framework in the event classification process. The second one (*Experiment #2*) aims at evaluating the performance of our proposed framework against an approach based on word embeddings, a more recent approach for contextual enrichment of tweets (Li et al., 2016)

With these two experiments, we aim at evaluating the contributions of:

- The semantic features to the classification process, compared to the use of textual features only;
- The pruning algorithm to the identification of relevant semantic features, possibly in combination with a feature selection algorithm;
- The use of conceptual features from external web documents, in addition to concepts extracted from tweets text only;
- The generality of the approach by considering events of distinct natures;
- The semantic enrichment compared to a word embeddings approach.

The experiments reported here are restricted to the pruning threshold strategy *QUARTILES* (Section 3.5) and the feature selection algorithm *CfsSubsetEval* (Hall, 1999), which overall yielded the best results. Additional results with other feature selection algorithms (e.g. InfoGain) and dataset preparations are reported in Romero (2017).

4.1. Target event datasets

We performed the experiments using seven target event of distinct natures. Table 1 presents the name of the datasets used as the target event, the number of tweets used from each dataset, the description and type of event it represents. Although some datasets originally contained more instances, we used a smaller number of instances to make the results comparable.

The positive events were extracted from the corresponding event dataset, and the negative ones were randomly selected from the other datasets. We also used a dataset of the SemEval-2016 Task 4, to simulate the existence of tweets not related to any particular event or subject. The final target event dataset has positive and negative examples in the approximate proportion of 1:2, since in a real-world crawling situation, we would collect much more negative examples than positive ones. For example, the Olympics dataset is composed of all posts from 2012 Olympic Games dataset as the positive label, and a random set of tweets belonging to the datasets FaCup, HSandy, Halloween, Alberta, Australia, Influenza and SemEval-2016 annotated as negative. To avoid depending on the same patterns used to crawl the data for event classification, we removed all keywords used for filtering the tweets.

4.2. Experiment #1: no contextual enrichment vs. hybrid semantic enrichment

In this first experiment, we compared the benefits of using the proposed hybrid semantic enrichment framework, against the classification of tweets based solely on textual features. By using a straightforward technique as the baseline, we can focus on the analysis of the contribution of our hybrid semantic enrichment approach in the classification of event-related tweets. We also analyzed the contribution of the different steps of the framework and how they collaborate to the improvement of the final results.

4.2.1. Experiment description

In this experiment, the baseline is composed of all alphabetic uni-grams extracted from each tweet dataset, including the normalized symbols (i.e. T_USER, T_EMOT, and T_URL). We compared this baseline with datasets prepared using the combination of all core features mentioned in Section 3.2. In other words, the dataset contains the uni-grams extracted from the tweets, incorporated with all semantic features resulting from the combination of agent, location, frequent terms and domain representative terms. Each dataset was prepared according to variations of the proposed framework to evaluate the contribution of the key stages, as follows:

- Experiment #1.1: Evaluate the contribution of the Semantic Feature Pruning step. Thus datasets are prepared according to all stages of the process, except for the Feature Selection step.
- Experiment #1.2: Evaluate the contribution of the Feature Selection step. To assess its complementary role to the Semantic Feature Pruning step, the datasets are prepared with and without the Semantic Feature Pruning step.
- Experiment #1.3: Evaluate the contribution of external document enrichment. Thus, data is pre-processed according to all steps of the hybrid enrichment framework, as well as without external enrichment.

As results can be dependent on the ability of the classification algorithm to handle highly dimensional datasets, we developed our analysis using three distinct classification algorithms widely used for tweet classification, namely Naive Bayes (NB), SVM (Support Vector Machine), and Random Forest (RF). NB is the simplest probabilistic classification algorithms (John & Langley, 1995), based on the application of the Bayes Theorem, which assumes total independence of variables. SVM is a linear supervised algorithm, which constructs a hyperplane that divides the space into dimensions representing classes, choosing the hyperplane that maximizes the distance from it to the nearest data point of each class. We adopted the Sequential Minimal Optimization (SMO) implementation of SVM (Platt, 1998). RF is an ensemble classification method

Table 2
Summarization of the experiments configuration.

Configuration	Experiment #1.1		Experiment #1.2		Experiment #1.3	
	Baseline	Enriched datasets	Baseline	Enriched datasets	Baseline	Enriched datasets
Pre-processing: Tokenization and normalization	X	X	X	X	X	X
Pre-processing: URL extraction		X		X		With and without
Conceptual feature extraction: From tweet text		X		X		X
Conceptual feature extraction: From web documents		X		X		With and without
Semantic enrichment		X		X		X
Semantic feature pruning		X		With and without		With and without
incorporation		X		X		X
Feature selection			X	X	X	X
Event classification	X	X	X	X	X	X
Resulting dataset configuration		Hybrid semantic enrichment		Hybrid semantic enrichment		Semantic-only, Hybrid semantic enrichment
Strategies for selecting features analyzed		QUARTILES	CFS	CFS, QUARTILES+CFS	CFS	CFS, QUARTILES+CFS

that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees (Breiman, 2001). All of these algorithms deal in different ways with high dimensionality.

We used the implementations for these algorithms available in Weka (Hall et al., 2009) (version 3.8.0), with default parameters, 10-fold cross-validation configuration, and 10 iterations. We compared the results using the F-Measure, Precision and Recall metrics. We also validate our results for each metric through a statistical test, using two-tail paired *t*-test, with a significance level of 0.05, using Weka Experimenter.

4.2.2. Dataset preparation

We prepared the enriched datasets according to the proposed framework (Fig. 1), but depending on the goal of the experiments, some of the steps were not executed. Table 2 summarizes this information.

Pre-processing: Performed as described in Section 3.2. To extract the content of referenced URLs, we employed Text Extraction of AlchemyAPI[2].

Conceptual feature extraction: For each target event (i.e. positive examples), we composed a list of agents, locations, frequent, and representative terms extracted from the tweets. When external document enrichment was deployed, a list of agents, locations, and representative terms were also extracted from the external web documents. We employed Open Calais API[3] to extract agents (i.e. person and organization), and locations (i.e. country, city, continent, and province). To extract the vocabulary features from tweet texts, we selected the top-20 frequent terms and defined a threshold of 15 for the TF-IDF selection (Romero & Becker, 2017). We applied only TF-IDF weighted to select the vocabulary from the web documents, using a threshold of 5, due to the volume of information presented in the web documents.

Semantic enrichment: We used the DBPedia and the *rdf:type* property. We employed the RapidMiner platform[4] (version 7.0), and two LOD Extension operators: (a) DBPedia Spotlight (to connect the conceptual features to the respective URI from DBPedia) and (b) Direct Type.

Semantic feature pruning: We applied the PageRank-based pruning method with the QUARTILES threshold strategy (Section 3.5).

Incorporation: Performed as described in Section 3.6.

Feature selection: We adopted the *CfsSubsetEval* implementation available in Weka platform, using *BestFirst* as the search method.

In Table 3, we present the amount of features resulting from each configuration: (a) textual features only (*TF*) for the baseline of each event type; (b) textual features incorporated with the semantic features without pruning (*WP*); (c) textual features incorporated with pruned semantic features (*QUARTILES*); (d) textual/semantic features selected by CfsSubsetEval only (*CFS*); and (e) textual and pruned semantic features selected by *CfsSubsetEval* (*QUARTILES+CFS*).

4.2.3. Experiment #1.1: semantic feature pruning step

In this section, we evaluate the contribution of the pruning algorithm for the performance of the semantic enrichment framework, comparing it against the baseline.

DBPedia provided good coverage for finding resources related to the conceptual features through RapidMiner LOD operators (80% in average). As it can be seen in Table 3, the semantic enrichment step considerably increases the number of features (94.2% in average). Using the pruning algorithm, this volume is significantly reduced (41% in average). A qualitative analysis of the remaining features was not performed since the number of remaining features after pruning was still too elevated.

We classified the TF, WP and QUARTILES datasets using three classification algorithms. Table 4 presents the Precision, Recall, and F-Measure results for the *Positive* class since we aim at identifying the tweets related to a specific event. Considering a significance level of $\alpha = 0.05$ results depicted with (*) represent that the baseline is statistically superior, whereas the (v) symbol means that the combination analyzed is statically superior against the baseline. Otherwise, there is no statistic difference between the results.

Discussions

Regarding the baseline, NB produced the best results in general. It is important to stress that the performance for some baseline cases was quite good, leaving room only for marginal improvements, particularly with regard to Precision (between 98% and 100%).

Considering the results of the proposed approach using pruning as the sole technique to select discriminant semantic features, we were able to improve the results in 44.4% of cases. An im-

[2] http://www.alchemyapi.com/products/alchemylanguage/text-extraction
[3] http://www.opencalais.com/
[4] http://rapidminer.com/

Table 3
Number of features resulting from the different steps of the framework.

Dataset	FaCup	Olympics	Halloween	HSandy	Alberta F.	Australia B.	Influenza
TF	1672	1825	1829	2127	1956	2092	1900
WP	2182	3723	4197	4311	5068	4055	2657
QUARTILES	1711	1971	2028	2349	2236	2309	2007
CFS	77	88	131	71	29	46	52
QUARTILES+CFS	70	101	146	70	29	58	53

Table 4
Statistical comparison between the baseline and the hybrid semantic enrichment configuration with pruning only.

Dataset	Algor.	Baseline			QUARTILES		
		P	R	F	P	R	F
FaCup	NB	0.936	0.808	0.867	**0.961 v**	0.784	0.863
	SMO	0.940	0.910	0.924	0.946	0.912	0.928
	RF	0.974	0.864	0.916	0.980	0.867	0.920
Olympics	NB	0.725	0.711	0.717	0.680 *	**0.783 v**	0.728
	SMO	0.881	0.824	0.851	0.883	0.821	0.850
	RF	0.967	0.670	0.791	0.964	0.644	0.771
Halloween	NB	0.858	0.733	0.790	0.831 *	0.725	0.774
	SMO	0.897	0.888	0.892	0.895	0.888	0.892
	RF	0.923	0.846	0.882	0.929	0.836	0.880
Hsandy	NB	0.916	0.849	0.881	0.885 *	0.830	0.856 *
	SMO	0.966	0.919	0.942	0.960	0.920	0.939
	RF	0.987	0.901	0.942	0.967 *	0.887	0.925 *
Alberta Floods	NB	0.947	0.952	0.949	**0.978 v**	**0.982 v**	**0.980 v**
	SMO	0.999	0.992	0.995	0.998	**1.000 v**	**0.999 v**
	RF	0.996	0.957	0.976	0.998	**0.989 v**	**0.994 v**
Australia Bushfire	NB	0.930	0.958	0.944	**0.980 v**	**0.988 v**	**0.984 v**
	SMO	0.999	0.992	0.995	0.997	0.996	0.997
	RF	0.994	0.984	0.989	0.983 *	0.992	0.987
Influenza	NB	0.961	0.998	0.979	**0.974 v**	0.995	0.984
	SMO	1.000	0.997	0.999	1.000	0.997	0.999
	RF	0.999	0.997	0.998	0.999	0.994	0.996

provement is given by the positive difference between the results achieved with our approach and the baseline, considering the same metric. For example, considering the Recall metric for the Olympics dataset, our approach yielded a better result (0.783) compared to the baseline (0.711) for the NB algorithm, with an improvement of 7.2 percentage points (pp). Improvements ranged from 0.1 pp to 7.2 pp.

Despite the good results in general, we observed that our approach statistically outperformed the baseline only in 20.6% of cases. The most improved metric was Recall, with 23.8% of statistically significant improvements, in general with no harm to Precision. It means that semantic enrichment improves the ability to recognize event-related tweets. The statistically significant improvements were associated in general to the NB classification algorithm and the Alberta F. target event. Actually, the impact depends on the ability of the algorithm to handle highly dimensional datasets.

In summary, the proposed hybrid semantic approach improves to some extent the ability to recognize event-related tweets. However, using pruning as the only means to reduce the number of semantic features produced modest improvements in the event classification performance.

4.2.4. Experiment #1.2: the feature selection step

In this section, we evaluate whether a feature selection technique could replace the proposed pruning method, or should be used in combination with one. We claim that they address complementary issues: whereas pruning selects the most representative semantic features according to the event domain at hand, the feature selection algorithm selects the most representative ones for the classification task. The latter thus assumes indistinctly semantic/textual features, as well as their contribution with regard to both positive and negative labels.

According to this evaluation goal, we adopted three datasets for each event: (a) a variation of the baseline to which the feature selection algorithm was also applied (Baseline+CFS), the CFS dataset (feature selection only), and the QUARTILES+CFS dataset, which combines feature selection and pruning. Table 3 displays the resulting number of features for the latter two.

Table 5 summarizes the results for the *Positive* class using the three classification algorithms. The same conventions for statistically significant results are adopted (v and * as in Table 4).

Discussions

a) Quantitative and qualitative analysis of selected features

Table 3 reports the number of features according to each feature reduction technique. Comparing to pruning only (QUARTILES), we can observe that the *CfsSubsetEval* method selects a significantly smaller number of features (CFS). However, when combined with pruning (QUARTILES+CFS), this number of features is not much different.

We manually analyzed the results to observe the characteristics of the features selected. Considering the CFS datasets, we observed that these datasets presented features highly related to the event target in the top of the resulting list. However, in general, few semantic features related to the event (positive class) were present (ranging from 2 to 17).

Considering the QUARTILES+CFS datasets, we observed a more balanced number of semantic and textual features (about 50% each). In addition, the selected semantic features better generalize the event domain, compared to CFS datasets. Thus, the pruning algorithm is key to boosting the value of semantic enrichment and does address the problem of selecting the proper generalization level of domain concepts retrieved from the LOD.

In conclusion, the combination of pruning with a general-purpose feature selection algorithm results in more discriminative textual and semantic features included in the dataset submitted to the classification algorithm.

b) Classification performance

Table 5 details the results of the hybrid semantic enrichment approach considering the application of a general-purpose feature selection algorithm only (CFS) and its combination with the pruning (QUARTILES+CFS) for the three classification algorithms.

The CFS datasets outperformed the baseline in 58.7% of cases, producing improvements that range from 0.1 pp to 7.6 pp in specific situations (i.e. Recall metric for the Olympics dataset, using the RF algorithm). However, these improvements were statistically significant in 36.5% of cases, mainly for the Alberta F. and Australia B. datasets, and the Recall metric.

More improvements were achieved with the QUARTILES+CFS combination (60.3% of cases). These improvements were statistically significant in 28.5% of cases. The Recall was the most affected metric, where 47.6% of the improvements were statistically significant, confirming the potential of the approach for recognizing events. The average improvement was about 3 pp, ranging from 0.1 pp to 32.5 pp in specific cases (i.e. Recall metric, Olympics dataset, and the NB algorithm). Recall that the baseline Precision left almost no room for improvement in many cases. QUARTILES+CFS

Table 5
Statistical comparison between the baseline and the hybrid semantic enrichment- *CFS* and *QUARTILES+CFS* variations.

Dataset	Algor.	Baseline+CFS			CFS			QUARTILES+CFS		
		P	R	F	P	R	F	P	R	F
FaCup	NB	0.967	0.730	0.832	0.941 *	0.707	0.807 *	0.926	**0.803 v**	**0.860 v**
	SMO	0.978	0.770	0.861	0.966	**0.822 v**	**0.888 v**	0.968	**0.815 v**	**0.885 v**
	RF	0.970	0.785	0.867	0.957	**0.835 v**	**0.891 v**	0.951	**0.832 v**	0.887
Olympics	NB	0.970	0.453	0.616	0.826 *	**0.519 v**	0.637	0.672 *	**0.778 v**	**0.705 v**
	SMO	0.956	0.611	0.744	0.937	**0.662 v**	0.775	0.952	0.622	0.751
	RF	0.971	0.572	0.718	0.923 *	**0.648 v**	**0.760 v**	0.968	0.580	0.725
Halloween	NB	0.805	0.827	0.816	0.808	**0.867 v**	0.836	0.809	0.847	0.827
	SMO	0.842	0.878	0.859	0.846	0.879	0.862	0.847	0.875	0.860
	RF	0.868	0.831	0.848	0.875	0.839	0.857	0.858	0.840	0.847
Hsandy	NB	0.976	0.822	0.892	0.912 *	**0.854 v**	0.882	0.908 *	0.848	0.877
	SMO	0.976	0.874	0.922	0.957 *	0.860	0.906 *	0.952 *	0.849	0.897 *
	RF	0.972	0.892	0.930	0.926 *	0.902	0.914 *	0.927 *	0.893	0.909 *
Alberta Floods	NB	0.991	0.971	0.981	0.996	**1.000 v**	**0.998 v**	0.998	**1.000 v**	**0.999 v**
	SMO	0.999	0.987	0.993	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
	RF	0.997	0.988	0.993	0.998	**1.000 v**	**0.989 v**	0.998	**1.000 v**	**0.999 v**
Australia Bushfire	NB	0.957	0.944	0.950	**0.982 v**	**0.994 v**	**0.988 v**	**0.980 v**	**0.987 v**	**0.983 v**
	SMO	0.992	0.981	0.986	0.992	**0.996 v**	**0.994 v**	0.985	**0.998 v**	0.992
	RF	0.992	0.986	0.989	0.996	**0.997 v**	**0.996 v**	0.985	**0.998 v**	0.991
Influenza	NB	0.999	0.997	0.998	0.999	0.995	0.997	0.999	0.998	0.998
	SMO	1.000	0.997	0.998	1.000	0.997	0.999	1.000	0.997	0.999
	RF	0.999	0.997	0.998	0.999	0.997	0.998	0.999	0.998	0.998

Table 6
Summarization of all results.

Configuration	General improvement	Statistically superior	Statistically inferior	Minimum improvement	Maximum improvement	Most improved metric	Most improved dataset
Without Pruning	30.1%	15.8%	39.6%	0.1 pp	4.5 pp	Recall	Alberta F.
CFS only	58.7%	36.5%	14.2%	0.1 pp	7.6 pp	Recall	FaCup, Olympics, Australia B., Alberta F.
QUARTILES only	44.4%	20.6%	11.1%	0.1 pp	7.2 pp	Recall	Alberta F.
QUARTILES+CFS	60.3%	28.5%	11.1%	0.1 pp	32.5 pp	Recall	FaCup, Australia B., Alberta F.

produced better results with regard to Precision, as more statistically inferior results were observed in CFS. The most affected datasets were FaCup, Alberta F., and Australia B. In general, the most robust classification to highly dimensional datasets was NB, considering the QUARTILES+CFS configurations.

c) Comparative analysis

In Table 6, we summarize the analyses regarding experiments *#1.1* and *#1.2*. For each variation, we analyzed the general improvement, the percentage of statistically superior results, the percentage of results that were statistically outperformed by the baseline, the minimum and the maximum improvements achieved. Considering the statistical analysis, we also highlight the most improved metric and the most improved target events[5]

In comparison to feature selection only, the results combining the pruning and the feature selection algorithms are slightly better. The latter produced better results overall and a significantly better maximum improvement. In addition, the baseline was statistically superior in fewer cases.

Regarding the metrics, in all cases, the largest number of improvements could be noticed for the Recall metric, though at the expense of Precision in some cases. In that sense, CFS produced more harm to the Precision metric. We can conclude that the hybrid semantic approach with adequately selected features presented good performance in recognizing the tweets related to the target event, and that the pruning algorithm is key for taking full benefit from the semantic enrichment process. The approach is suitable for datasets representing events of distinct natures, and thus it is generalizable.

[5] We considered that the dataset must have at least half the number of cases statistically superior to the baseline than the dataset that presented the highest gain, considering all classification algorithms. For example, in the QUARTILES+CFS combination, the Australia Bushfire dataset statistically outperformed the baseline in seven cases for the CFS strategy. Thus, to be inserted in this table, the others datasets must have, at least, 3.5 cases that statistically outperformed the baseline, considering all classification algorithms.

Table 7
Amount of textual and semantic features for each configuration.

Dataset	Contextual Enrich.	WP	CFS	QUARTILES+CFS
FaCup	SOE	2100	64	61
	HSE	2182	77	70
Olympics	SOE	2417	90	95
	HSE	3723	88	101
Halloween	SOE	2983	131	139
	HSE	4197	131	146
HSandy	SOE	2577	62	70
	HSE	4311	71	70
Alberta Floods	SOE	2552	50	52
	HSE	5068	29	29
Australia Bushfire	SOE	2785	44	47
	HSE	4055	46	58
Influenza	SOE	2371	53	55
	HSE	2657	52	53

4.2.5. Experiment #1.3: semantic-only vs. hybrid semantic enrichment

Our final comparison is between the hybrid semantic enrichment (tweets and related web documents) and the semantic-only enrichment (tweets only). We develop our comparison considering the CFS and QUARTILES+CFS dataset configurations.

In Table 7, we present the resulting number of features of each dataset configuration according to the contextual enrichment technique: *semantic-only enrichment* (*SOE*) and *hybrid semantic enrich-*

Table 8
Statistical comparison between the baseline and the semantic-only enrichment configuration.

Dataset	Algor.	Baseline+CFS			CFS			QUARTILES+CFS		
		P	R	F	P	R	F	P	R	F
FaCup	NB	0.967	0.730	0.832	0.917 *	**0.791 v**	0.849	0.970	0.762	0.854
	SMO	0.978	0.770	0.861	0.951 *	**0.822 v**	0.881	0.972	**0.802 v**	0.878
	RF	0.970	0.785	0.867	0.945 *	**0.828 v**	0.882	0.967	0.812	0.882
Olympics	NB	0.970	0.453	0.616	0.721 *	**0.544 v**	0.619	0.744 *	**0.525 v**	0.615
	SMO	0.956	0.611	0.744	0.921 *	0.635	0.751	0.944	0.612	0.741
	RF	0.971	0.572	0.718	0.907 *	**0.636 v**	0.746	0.929 *	0.584	0.716
Halloween	NB	0.805	0.827	0.816	0.826	0.836	0.830	0.801	0.820	0.810
	SMO	0.842	0.878	0.859	0.838	0.880	0.858	0.842	0.876	0.858
	RF	0.868	0.831	0.848	0.856	0.856	0.856	0.864	0.831	0.847
Hsandy	NB	0.976	0.822	0.892	**0.988 v**	**0.883 v**	**0.932 v**	0.916 *	**0.880 v**	0.898
	SMO	0.976	0.874	0.922	0.986	**0.920 v**	**0.952 v**	0.935 *	**0.918 v**	0.926
	RF	0.973	0.892	0.930	**0.993 v**	**0.917 v**	**0.953 v**	0.941 *	**0.918 v**	0.929
Alberta Floods	NB	0.991	0.971	0.981	0.882 *	**1.000 v**	0.937 *	0.990	0.972	0.981
	SMO	0.999	0.987	0.993	1.000	0.989	0.994	0.999	0.987	0.993
	RF	0.997	0.988	0.993	0.998	0.988	0.993	0.997	0.989	0.993
Australia Bushfire	NB	0.957	0.944	0.950	0.897 *	**0.998 v**	0.945	0.841 *	**0.999 v**	0.913 *
	SMO	0.992	0.981	0.986	0.998	**0.997 v**	**0.998 v**	**0.997 v**	0.989	0.993
	RF	0.992	0.986	0.989	**0.998 v**	**0.997 v**	**0.997 v**	0.996	0.991	0.993
Influenza	NB	0.999	0.997	0.998	0.999	0.998	0.998	0.999	0.998	0.998
	SMO	1.000	0.997	0.998	1.000	0.997	0.999	1.000	0.997	0.998
	RF	0.999	0.997	0.998	1.000	0.997	0.999	0.999	0.997	0.998

ment (*HSE*). Table 8 shows the classification performance results for the Positive class using the semantic-only enrichment, using the same conventions for statistical significance. The results for the hybrid semantic enrichment configuration were presented in Table 5.

Discussions

a) Qualitative and quantitative analysis of selected features

In general, the adoption of external documents increased the number of named entities found. We observed that Sports-related tweets presented more URLs associated with others tweets, whilst the ones related to natural disasters and epidemics were linked to web documents with more relevant content (e.g. prevention measures, affected areas, and detailed reports). This difference is directly related to the difference of features reported in Table 7 (WP) when external documents are used (HSE). Some URLs were no longer valid and thus were ignored by the analysis, which might have influenced our results, for in these cases, fewer additional conceptual features were added.

We can also notice that the number of features is drastically reduced for all target events when some technique for selecting discriminative features is applied (CFS and QUARTILES+CFS). However, the final number of features is quite similar, regardless of the enrichment strategy.

We performed a manual analysis of selected features. In general, the hybrid enrichment strategy resulted in more semantic features in the final dataset, compared to the semantic-only strategy. Considering specifically the *CfsSubsetEval* algorithm (CFS), in both cases the textual and semantic features on the top of the list were very related to the event analyzed. Concerning to the combined application of pruning and feature selection (QUARTILES+CFS), we observed that except for two datasets (FaCup and Olympics), the resulting set of features for each enrichment configuration was very different. The hybrid semantic enrichment was the strategy that resulted in more meaningful semantic features.

We conclude that the adoption of external sources does increase the opportunity of finding domain-related concepts that are relevant to the classification task, which otherwise would not be identified in tweet contents.

b) Classification performance of semantic-only enrichment

Considering the results in Table 8, we also observed improvements in classification results when semantic enrichment is performed over tweet contents only. For the CFS datasets, there was an improvement in 68.2% of cases. For the QUARTILES+CFS datasets, there was an improvement in 39.6% of cases. The statistically significant improvements corresponded to 31.7% and 11.1% of cases, respectively. Improvements ranged from 0.1 pp to 9.1 pp. As in the previous experiments, the best results were observed for the Recall metric, but often at the expense of Precision.

In summary, the results were very similar in both configurations, and the application of the CFS enabled to achieve more statistically significant results. It is justified as the number of features resulting from semantic enrichment steps is much smaller. Thus, the general-purpose feature selection algorithm is enough to achieve good results. However, feature selection-only degraded the Precision more often, compared to QUARTILES+CFS.

c) Comparison between hybrid semantic and semantic-only enrichment configurations

To compare the boost produced by using external documents in the enrichment step, we used the semantic-only enrichment configuration as the baseline, and calculated the difference between them. Results are displayed in the boxplot of Fig. 4 for each metric.

Two main improvements can be noticed when comparing hybrid and semantic-only enrichment. For the CFS datasets, it improves the Precision, which was a metric degraded in some datasets. As for the combination of pruning and feature selection (QUARTILES+CFS), the main improvement lies in the Recall metric, where most results range between 0.9 and 1.3 pp, plus a few outliers. The median, lower quartiles and upper quartiles are superior, compared to CFS only. Overall, external documents balance better the Precision and Recall metrics in the hybrid enrichment, and thus F-Measure median, quartiles, and upper outliers are slightly superior, compared to CFS. Thus, the hybrid semantic approach does contribute to the identification of event-related tweets, by the augmentation of additional domain concepts.

As for the most affected algorithms, RF yielded the best improvements with semantic-only enrichment, whereas for the hybrid enrichment it was NB, followed by RF. Thus, we confirm previous findings (Do, Lenca, Lallich, & Pham, 2009; Statnikov, Wang, & Aliferis, 2008) that SMO is more robust to irrelevant features, and thus the reduction of features affects only to some extent its own internal feature selection method used for finding the separating

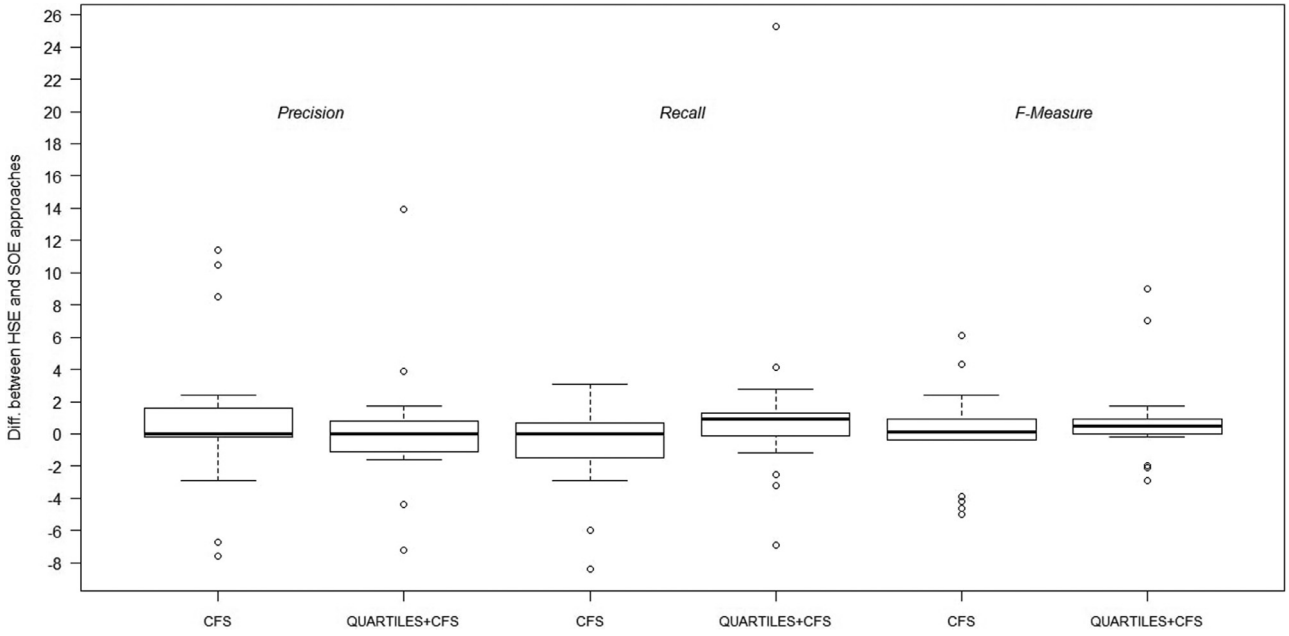

Fig. 4. Difference between the hybrid semantic and semantic-only enrichment configurations.

Table 9
Summarization of the results for semantic-only (SOE) and hybrid semantic enrichment (HSE), for the *CfsSubsetEval* algorithm and its combination with pruning.

Configuration	Type of Enrich.	General improvement	Statistically superior	Statistically inferior	Minimum improvement	Maximum improvement	Most improved metric	Most improved dataset
Without Pruning	SOE	31.7%	6.3%	42.8%	0.1 pp	4.0 pp	Recall	Alberta F., Australia B.
	HSE	30.1%	15.8%	39.6%	0.1 pp	4.5 pp	Recall	Alberta F.
CFS only	SOE	68.2%	31.7%	14.2%	0.1 pp	9.1 pp	Recall	HSandy, Australia B.
	HSE	58.7%	36.5%	14.2%	0.1 pp	7.6 pp	Recall	FaCup, Olympics, Australia B., Alberta F.
QUARTILES+CFS	SOE	39.6%	11.1%	11.1%	0.1 pp	7.2 pp	Recall	HSandy, Australia B.
	HSE	60.3%	28.5%	11.1%	0.1 pp	32.5 pp	Recall	FaCup, Australia B., Alberta F.

hyperplanes. RF and NB do deal with high dimensionality, but are sensitive to the dependencies among features. The CFS algorithm does solve to some extent this issue. It is confirmed by additional experiments using InfoGain (Romero, 2017), which yielded poor results (not reported here), showing unquestionably the value added by the pruning method.

In summary, semantic enrichment does contribute to better event-related tweet classification. Considering the results, we conclude that the *semantic-only enrichment* approach could be employed in situations where it is known that the dataset presents few URLs, or for old tweets in which URLs are no longer available. Otherwise, the *hybrid semantic enrichment* should be used, since it presented more statistically significant results. The more semantic features result from the semantic enrichment step, the bigger the value added by the pruning algorithm as a complementary technique for the selection of discriminative features is.

d) Final comparative analysis

Table 9 summarizes the experiments performed for the two types of enrichment. According to the information presented in rows Without Pruning, both strategies yielded quite poor results, confirming that techniques that select the proper discriminatory features are essential to semantic enrichment.

With regard to semantic-only enrichment for the CFS configuration, despite a greater number of improvements were observed, compared to the hybrid, this difference is compensated when evaluating their statistical significance. More importantly for our purposes, the hybrid enrichment enabled more generalized results, as it excelled in both sports-related and natural disaster events, which are quite different in structure and contents.

Considering the combination of pruning and feature selection, the superiority of the hybrid approach is unquestionable compared to semantic-only enrichment. It improved the results in 60.3%, with a maximum improvement of 32.5 pp (Recall) compared to the baseline. Also, it excelled in datasets of distinct natures. Regarding the Olympics dataset, the improvements using NB algorithm were expressive, but when using RF and SMO, the difference was not always statistically significant. Thus, Olympics did not fit the adopted criteria, despite the good results.

Regarding the classification algorithms, most statistically superior results were achieved using the NB and RF algorithms, for both configurations. The most improved metric was Recall, for all situations. The Alberta F. and Australia B. datasets also presented good results in all cases, followed by the HSandy dataset. The sportive datasets were presented in one and two situations, respectively, both using the *hybrid semantic enrichment* configuration.

In summary, the use of external information made it possible to obtain better results in more datasets, thus proving the generalization capacity of the proposed approach with regard to different types of events. Feature selection and pruning address different

Table 10
Comparison between the event classification in tweets using word embeddings against the hybrid semantic enrichment framework.

Dataset	Algor.	Word Embeddings			QUARTILES+CFS		
		P	R	F	P	R	F
FaCup	NB	0.891	0.666	0.831	0.915	0.914	0.912
	SMO	0.961	0.830	0.920	0.933	0.931	0.929
	RF	0.939	0.758	0.879	0.931	0.930	0.929
Olympics	NB	0.447	0.835	0.663	0.851	0.827	0.833
	SMO	0.949	0.575	0.820	0.899	0.893	0.884
	RF	0.821	0.458	0.743	0.896	0.886	0.875
Halloween	NB	0.704	0.607	0.743	0.883	0.881	0.882
	SMO	0.867	0.612	0.799	0.905	0.904	0.905
	RF	0.809	0.589	0.774	0.901	0.901	0.901
HSandy	NB	0.634	0.881	0.774	0.920	0.920	0.919
	SMO	0.906	0.735	0.864	0.936	0.935	0.934
	RF	0.877	0.651	0.827	0.939	0.940	0.939
Alberta Floods	NB	0.781	0.778	0.831	0.999	0.999	0.999
	SMO	0.891	0.747	0.862	0.999	0.999	0.999
	RF	0.865	0.678	0.831	0.999	0.999	0.999
Australia Bushfire	NB	0.733	0.851	0.824	0.989	0.989	0.989
	SMO	0.833	0.706	0.827	0.995	0.995	0.995
	RF	0.840	0.646	0.800	0.994	0.994	0.994
Influenza	NB	0.784	0.948	0.887	0.999	0.999	0.999
	SMO	0.983	0.959	0.978	0.999	0.999	0.999
	RF	0.980	0.865	0.938	0.999	0.999	0.999

issues concerning to the identification of discriminatory features, and the results variation favoring one or other reduction strategy is dependent on the robustness of the algorithm applied. In general, SMO is less sensitive, *CfsSubsetEval* affects more RF due to its ability to find redundant features, and pruning combined with *CfsSubsetEval* plays a more important role about NB.

4.3. Experiment #2: hybrid semantic enrichment vs. word embeddings

This second experiment aims at evaluating the performance of the proposed framework against an alternative form of enrichment, based on word embeddings. We choose this approach since it captures the syntactic and semantic characteristics of a word, allowing their representation in different contexts. Thus, it allows us to compare our hybrid enrichment approach (i.e. QUARTILES+CFS configuration) to another alternative for contextual enrichment.

4.3.1. Building the baseline

Word embeddings is a distributional semantic approach, which produces word vectors for each word in the vocabulary. We employed the pre-trained word vectors using GloVe[6], produced over 2 billion tweets, representing a 1.2 million vocabulary. We employed the GloVe model built in a 100-dimension space.

To combine these word embeddings with the tweets of the target datasets, we employed the mean of the individual term's vector (Liu, Liu, Chua, & Sun, 2015). Specifically, for each word in the tweet, we search for that word in the word embeddings model. If the corresponding word exists in the model, we store its word vector. Then, we calculate the mean of all word vectors found. This aggregation allows a condensed embedding-based features representation, in which each tweet is represented by a unique vector, containing a 100-dimensional array.

We used the Gensim[7] and Scikit Learn Python libraries. For the classification, we employed the implementations available for NB, SMO, and RF.

[6] http://nlp.stanford.edu/projects/glove/
[7] https://github.com/RaRe-Technologies/gensim

4.3.2. Results

In Table 10, we present the results of the event classification task according to the word embeddings approach (baseline) and our hybrid semantic enrichment framework. These results represent one iteration of a 10-fold cross-validation configuration, considering the weighted Precision, Recall, and F-Measure metrics, respectively. The average differences between our approach and the baseline per metric and algorithm are detailed in Fig. 5.

We validate our results through a statistical test, using two-tail paired *t*-test, with a significance level of 0.05. For the comparison, we analyzed a group of results (i.e. each dataset variation and classifier against the corresponding baseline built using word embeddings), using the Microsoft Excel.

4.3.3. Discussions

a) Word embeddings for event classification in tweets

In Table 10, we observe that the word embeddings approach produces good results for the event classification task, but not for all target events tested. The results using the SMO algorithm were slightly better than the ones produced by the other algorithms. Good results could be noticed for the Precision metric, mainly for the SMO and RF algorithms. For NB, better results could be noticed in terms of Recall. This difference in performance affects the comparative results discussed in this section.

In comparing the baseline with the QUARTILES+CFS datasets, we could observe improvements in 93.6% of cases, ranging from 0.9 pp to 42.8 pp in specific situations (i.e. the Olympics dataset, considering the Recall metric and RF algorithm). The average improvement considering all classification algorithms was about 11.36 pp, 21.23 pp, and 11.42 pp, for Precision, Recall, and F-Measure metrics, respectively. The baseline was superior only in a few cases (Precision metric for the sportive datasets, with the SMO classifier).

b) Classification performance comparison

In Fig. 5, we compare the difference between our approach and the word embeddings baseline. For the NB classifier, the difference was higher in terms of Precision, but for SMO and RF, the difference was higher for Recall. As mentioned, this difference is explained by the results produced by each algorithm over the baseline. The median improvements considering all metrics were 16.8 pp, 10.1 pp, and 13.4 pp, for NB, SMO, and RF respectively.

The superior results achieved by our approach were statistically significant in 88.8% of cases. The results of all classifiers for Recall and F-Measure metrics presented statistically significance difference against the baseline, while for the Precision metric, this difference could be noticed for NB and RF classifiers.

In summary, these results show that our solution is a feasible and generalizable contextual enrichment method to support the classification of distinct event types. The solution was robust to three distinct algorithms widely used for text classification and outperformed the results achieved using a word embeddings approach, which has been increasingly adopted for contextual enrichment in different applications.

5. Conclusion and future work

In this work, we proposed a hybrid semantic enrichment framework to improve the event-related classification of tweets. The approach combines semantic enrichment with two other contextual enrichment strategies, namely external document enrichment and named entity extraction. Each one of them has a specific role in providing context to the poor and sparse content of tweets, and help in the event classification task. We also addressed how to select discriminative features using two complementary techniques: a specific-purpose pruning algorithm and a general-purpose feature selection algorithm. These elements were evaluated in a broad

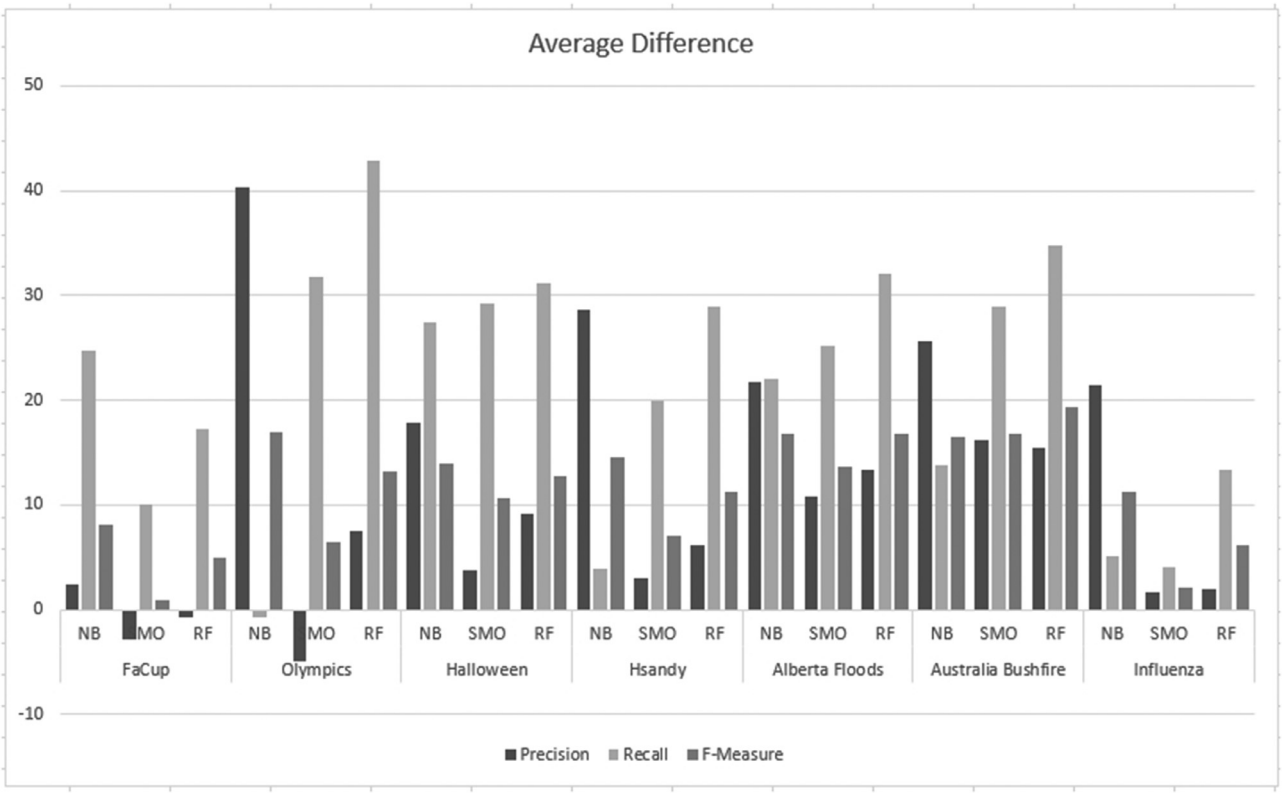

Fig. 5. Difference between the Hybrid Semantic Enrichment and the Word Embeddings approach, for all classifiers.

experimental setting. The proposed approach does not rely on assumptions about the event type, and thus it can be applied to a broad range of events, the results can be compared to each other, as well as be used as the baseline for future event-related tweet classification approaches.

Each enrichment technique integrated plays a different role in improving event classification. NER tools provide categorical context to vocabulary, which complements frequent and representative terms (Romero & Becker, 2016; 2017). External enrichment helps to overcome the noisy and sparse nature of tweets, particularly in the typical situation in which tweets do tend to reference valid URLs. Semantic enrichment relates the content of documents to domain concepts, but to be effective, techniques to find discriminative features among the huge volume of resulting features are required.

We proposed a pruning method that handles the proper generalization level with regard to the event at hand. It is complementary to the role of feature selection methods: it helps to select the semantic features that are relevant to the domain, whereas the ones that improve the classification. The feature selection method assessed in this paper (i.e. *CfsSubsetEval*) deals with the relevance with regard to the target class, and redundancy among features. The results favored the isolated or the combined approach according to the classification algorithm. In Romero and Becker (2017), we assessed the *InfoGain* algorithm and the role of the pruning was outstanding.

In general, the results show that the proposed hybrid semantic enrichment framework is a feasible and generalizable solution to support the classification of distinct event types, where the extent of the improvement depends on the target event and the algorithm. Considering the textual features baseline, it achieved improvements in 60.3% of cases, whereas the improvements could be noticed in 93.6% of cases for the baseline using word embeddings. The main statistically significant improvement lies in the Recall metric and, overall, external documents balance better the Precision and Recall metrics in the hybrid enrichment. Despite the promising results in datasets representing events of a distinct nature, no patterns could be found with regard to improvements in all examples of a specific event type (e.g. sportive events - FaCup and Olympics). The assessment of the approach using additional target events, and a higher volume of tweets is a means to further confirm the current results.

Considering related work, our major strengths are: (a) we consider a broad range of events since we do not rely on assumptions about the event types, different from related work (Abel et al., 2012; Sakaki et al., 2010; Schulz et al., 2013; Singh et al., 2017), and we defined core features to be extracted from the documents; (b) we deal with the huge number of features resulting from semantic and external documents enrichment, a characteristic addressed only by Khare et al. (2018) using feature selection; (c) unlike other works that adopted hybrid semantic enrichment (Abel et al., 2012; Packer et al., 2012), we do not rely on the existence of prior information; and (d) we assessed the contribution of each framework component, rather than the performance of a complete solution. As limitations, related work that used semantic enrichment (Khare et al., 2018; Rowe & Stankovic, 2011; Schulz et al., 2015; Schulz et al., 2013; Tonon et al., 2017) proposed other properties and knowledge bases. However, the knowledge base and semantic property chosen (DBPedia and *rdf:type*) was a common characteristic in these works (Khare et al., 2018; Rowe & Stankovic, 2011; Schulz et al., 2013). In terms of scope, Chen et al. (2018), Sakaki et al. (2010) and Liu et al. (2016) also addressed the event identification task.

As future work, we plan to address some of the limitations of the present work. First, we consider only the event classification task, without relating it to the event identification task, which fo-

cuses on creating groups of subject and time-related tweets from large data streams. To address this limitation, we need to integrate the hybrid semantic enrichment framework into an architecture that deals with the clustering of topic-related tweets to encompass both tasks. Second, the results show that the proposed approach is a feasible and generalizable solution to support the classification of distinct event types. However, our method may not be as effective when compared to specific-domain approaches. Third, we tested our approach with datasets of similar sizes in order to compare results. The behavior considering large datasets must also be assessed, particularly with regard to a word embeddings baseline. Fourth, the resulting event classification models reflect the training data. Hence, another line of work is to investigate the incremental construction of models in order to capture concept drift and to adapt an existing model to more generic or specific types of event in the same domain (e.g. a specific soccer game vs. sportive events). We could also extend the assessment of the proposed framework experimenting with: (a) other properties and knowledge bases available in the LOD cloud; (b) alternative feature selection techniques and classification algorithms; (c) an approach to cluster similar events that occur in different places and periods; and (d) more event datasets, as well as datasets of different sizes.

Acknowledgments

We thank Valentina Sintsova for making the Olympics dataset available. This research was financially supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) (1486104) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) (459322/2014-1) - Brazil.

Appendix A. Related work summary

Table A.11
Summary of related work.

Work	Features			Selection of features	Learning technique	Event type
	Tweet	External	Semantic			
Sakaki et al. (2010)	keywords and context words	NO	NO	NO	SVM	Natural disasters
Vosecky et al. (2014)	Named entities and timestamps	URL contents	NO	NO	K-means, DBSCAN, Single-pass Incremental Clusterer, and Direct	General (volume)
Chen et al. (2018)	Textual content, hashtags, time and location	Retweet contents	NO	NO	RL-LDA	General (retweet behavior)
Sicilia et al. (2018)	Statistical features, sentiment analysis, URL	Retweet	NO	YES	SVM, RF, AdaBoost, Random Tree, Nearest Neighbour	Health
Rosa et al. (2011)	Hashtags and TF-IDF	URL contents	NO	NO	LDA, K-Means, and Rocchio	General (popular hashtags)
Genc et al. (2011)	BOW	Wikipedia	NO	NO	LSA	General (similarity)
Wang and Niu (2018)	Unspecified	Climate documents	NO	NO	Clusterer	Climate events
Rowe and Stankovic (2011)	Named entities	NO	DBPedia (dct:subject, rdf:type)	YES	Proximity-based clustering and NB	General (volume)
Schulz et al. (2013)	TF-IDF, temporal/spatial expr., named entities	NO	DBPedia (rdf:subClassOf, skos:broader)	NO	SVM, NB, and JRip	Incidents
Schulz et al. (2015)	Named entities, temporal/spatial expr.	NO	DBPedia (rdf:type, dct:subject)	NO	J48, NB, JRip, SVM, RF	Incidents
Tonon et al. (2017)	Named entities	NO	DBPedia, Wordnet	NO	NA	Emergency situations
Khare et al. (2018)	Statistical features	NO	DBPedia, BableNet (rdf:type, rdfs:label, dct:subject, dbo:country)	YES	SVM	Emergency situations
Edouard et al. (2017)	Named entities	NO	DBPedia, YAGO	NO	NB, SVM, LSTM	General (volume)
Packer et al. (2012)	Named entities	Concert program	YAGO2	NO	Pearson's Correlation	Rock concert
Abel et al. (2012)	Named entities	Broadcasting service and URL contents	DBPedia (rdfs:label)	NO	Hand-crafted rules	Emergency situations
Li et al. (2016)	Unspecified	Google News and other general data collections	Socialbaker	NO	SVM	General (news)

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). Twitcident: Fighting fire with information from social web streams. In *Proceedings of the 21st international conference on world wide web WWW '12 Companion* (pp. 305–308). New York, NY, USA: ACM. doi:10.1145/2187980.2188035.

Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., & Gomes, J. B. (2016). A rule dynamics approach to event detection in twitter with its application to sports and politics. *Expert Systems with Applications, 55*, 351–360.

Anantharam, P., Barnaghi, P., Thirunarayan, K., & Sheth, A. (2015). Extracting city traffic events from social streams. *ACM Transactions on Intelligent Systems and Technology*, (4), 43:1–43:27. doi:10.1145/2717317.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, (1), 132–164. doi:10.1111/coin.12017.

Becker, H., Iter, D., Naaman, M., & Gravano, L. (2012). Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on web search and data mining WSDM '12* (pp. 533–542). New York, NY, USA: ACM. doi:10.1145/2124295.2124360.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. *Fifth AAAI conference on weblogs and social media*.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, X., Zhou, X., Sellis, T., & Li, X. (2018). Social event detection with retweeting behavior correlation. *Expert Systems with Applications, 114*, 516–523. doi:10.1016/j.eswa.2018.08.022.

Do, T.-N., Lenca, P., Lallich, S., & Pham, N.-K. (2009). Classifying very-high-dimensional data with random forests of oblique decision trees,.

Edouard, A., Cabrio, E., Tonelli, S., & Le Thanh, N. (2017). Semantic linking for event-based classification of tweets. *International Journal of Computational Linguistics and Applications*, 12.

Fisichella, M., Stewart, A., Cuzzocrea, A., & Denecke, K. (2011). Detecting health events on the social web to enable epidemic intelligence. In *Proceedings of spire 2011: 18th edition of the international symposium on string processing and information retrieval*.

Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). *Discovering context: Classifying tweets through a semantic transform based on wikipedia* (pp. 484–492). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. *Technical Report*.

Janpuangtong, S., & Shell, D. A. (2015). Leveraging ontologies to improve model generalization automatically with online data sources. In B. Bonet, & S. Koenig (Eds.), *Proceedings of the twenty-ninth AAAI conference on artificial intelligence, january 25–30, 2015, Austin, Texas, USA* (pp. 3981–3986). AAAI Press.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Eleventh conference on uncertainty in artificial intelligence* (pp. 338–345). San Mateo: Morgan Kaufmann.

Khare, P., Burel, G., & Alani, H. (2018). Classifying crises-information relevancy with semantics. In *The semantic web - 15th international conference, ESWC 2018, Heraklion, Crete, Greece, june 3–7, 2018, proceedings* (pp. 367–383). doi:10.1007/978-3-319-93417-4_24.

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. *NAACL*.

Li, Q., Nourbakhsh, A., Shah, S., & Liu, X. (2017). Real-time novel event detection from social media. In *33rd IEEE international conference on data engineering, ICDE 2017, San Diego, CA, USA, april 19–22, 2017* (pp. 1129–1139). doi:10.1109/ICDE.2017.157.

Li, Q., Shah, S., Liu, X., Nourbakhsh, A., & Fang, R. (2016). Tweet topic classification using distributed language representations. In *Proceedings of the 2016 IEEE/WIC/ACM international conference on web intelligence, Omaha, Nebraska, USA*.

Liu, X., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., et al. (2016). Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. In *Proceedings of the 25th international conference on information and knowledge management CIKM '16* (pp. 207–216). New York, NY, USA: ACM.

Liu, X., Nourbakhsh, A., Li, Q., Shah, S., Martin, R., & Duprey, J. (2017). Reuters tracer: Toward automated news production using large scale social media data. In *2017 IEEE international conference on big data (big data)* (pp. 1483–1493). doi:10.1109/BigData.2017.8258082.

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical word embeddings. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence AAAI'15* (pp. 2418–2424). AAAI Press.

Medvet, E., & Bartoli, A. (2012). Brand-related events detection, classification and summarization on twitter. In *Proceedings of the 2012 IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technology - volume 01 WI-IAT '12* (pp. 297–302). Washington, DC, USA: IEEE Computer Society.

Packer, H. S., Samangooei, S., Hare, J. S., Gibbins, N., & Lewis, P. H. (2012). Event detection using twitter and structured semantic query expansion. In *Proceedings of the 1st international workshop on multimodal crowd sensing CrowdSens '12* (pp. 7–14). New York, NY, USA: ACM. doi:10.1145/2390034.2390039.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. *Technical Report, 1999-66*. Stanford InfoLab.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods - support vector learning*. MIT Press.

Ristoski, P., & Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web, 36*(Supplement C), 1–22.

Romero, S. (2017). *A framework for event classification in tweets based on hybrid semantic enrichment*. Brazil: Pós-Graduação em Ciência da Computação - Universidade Federal do Rio Grande do Sul Master's thesis.

Romero, S., & Becker, K. (2016). Experiments with semantic enrichment for event classification in tweets. In *Proceedings of the 2016 IEEE/WIC/ACM international conference on web intelligence, Omaha, Nebraska, USA* (pp. 503–506). doi:10.1109/WI.2016.83.

Romero, S., & Becker, K. (2017). Improving the classification of events in tweets using semantic enrichment. In *Proceedings of the international conference on web intelligence WI '17* (pp. 581–588). New York, NY, USA: ACM. doi:10.1145/3106426.3106435.

Rosa, K. D., Shah, R., Lin, B., Gershman, A., & Frederking, R. (2011). Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*.

Rowe, M., & Stankovic, M. (2011). Aligning tweets with events: Automation via semantics. *Semantic Web Journal*.

Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, & M. Hauswirth, et al. (Eds.), *CEUR workshop proceedings lecture notes in computer science* (pp. 56–66). Springer Berlin Heidelberg. doi:10.1007/978-3-642-35176-1_32.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web WWW '10* (pp. 851–860). New York, NY, USA: ACM. doi:10.1145/1772690.1772777.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009). TwitterStand: News in Tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems GIS '09* (pp. 42–51). New York, NY, USA: ACM. doi:10.1145/1653771.1653781.

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, & D. Vrandečić, et al. (Eds.), *The semantic web – ISWC 2014 se - 16 lecture notes in computer science* (pp. 245–260). Springer International Publishing. doi:10.1007/978-3-319-11964-9_16.

Schulz, A., Guckelsberger, C., & Janssen, F. (2015). Semantic abstraction for generalization of tweet classification: An evaluation on incident-related tweets. *Semantic Web Journal*, 1–21. doi:10.3233/SW-150188.

Schulz, A., Ristoski, P., & Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. In *ESWC 2013: vol. 7955* (pp. 22–33). Springer Berlin Heidelberg. Lecture Notes in Computer Science. doi: 10.1007/978-3-642-41242-4_3.

Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Systems with Applications, 110*, 33–40. doi:10.1016/j.eswa.2018.05.019.

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2017). Event classification and location prediction from tweets during disasters. *Annals of Operations Research*. doi:10.1007/s10479-017-2522-3.

Sintsova, V., Musat, C., & Pu, P. (2013). Fine-grained emotion recognition in olympic tweets based on human computation. *4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics, 9*(1), 319.

Tonon, A., Cudré-Mauroux, P., Blarer, A., Lenders, V., & Motik, B. (2017). *Armatweet: Detecting events by semantic tweet analysis* (pp. 138–153)). Cham: Springer International Publishing.

Vosecky, J., Jiang, D., Leung, K. W.-T., Xing, K., & Ng, W. (2014). Integrating social and auxiliary semantics for multifaceted topic modeling in twitter. *ACM Transactions on Internet Technology*, (4), 27:1–27:24. doi:10.1145/2651403.

Wang, H., & Niu, Z. (2018). Climate event detection algorithm based on climate category word embedding. In *Proceedings of the 2018 international conference on big data and computing ICBDC '18* (pp. 71–77). New York, NY, USA: ACM. doi:10.1145/3220199.3220203.