UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

SIMONE APARECIDA PINTO ROMERO

A Framework for Event Classification in Tweets Based on Hybrid Semantic Enrichment

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof[a]. Dr[a]. Karin Becker

Porto Alegre
March 2017

CIP — CATALOGING-IN-PUBLICATION

Romero, Simone Aparecida Pinto

A Framework for Event Classification in Tweets Based on Hybrid Semantic Enrichment / Simone Aparecida Pinto Romero. – Porto Alegre: PPGC da UFRGS, 2017.

128 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2017. Advisor: Karin Becker.

1. Semantic Web. 2. DBPedia. 3. LOD. 4. Twitter. 5. Event Classification. I. Becker, Karin. II. Título.

“Put your heart, mind, and soul into even your smallest acts.
This is the secret of success.”

— Swami Sivananda

ACKNOWLEDGEMENTS

I thank God first for all the opportunities and for always guiding me on this long journey, helping me overcome each obstacle, allowing everything to work out in the end.

My sincere gratitude to my parents Maria and Hector (in memoriam), who always encouraged me and did the possible and the impossible for me to achieve my life goals. To Jhonny Mertz, for all love, affection, patience, and unconditional support in this journey.

I would like to express my special thanks to my dear friends Fernanda Brandão and Luiz Fernando de Souza for their support, affection, for the moments of joy and fun, and especially for the understanding of my absence in those two years.

I would also like to thank my colleagues from the lab 213, for their help and support.

Special thanks to my advisor Prof. Karin Becker, for her guidance for the development of this research, support, motivation, and for all the teachings.

Thanks to the financial assistance of CAPES and the infrastructure provided by the Federal University of Rio Grande do Sul, which allowed the development of this research.

ABSTRACT

Social Media platforms have become key as a means of spreading information, opinions or awareness about real-world events. Twitter stands out due to the huge volume of messages about all sorts of topics posted every day. Such messages are an important source of useful information about events, presenting many useful applications (e.g. the detection of breaking news, real-time awareness, updates about events). However, text classification on Twitter is by no means a trivial task that can be handled by conventional Natural Language Processing techniques. In addition, there is no consensus about the definition of which kind of tasks are executed in the Event Identification and Classification in tweets, since existing approaches often focus on specific types of events, based on specific assumptions, which makes it difficult to reproduce and compare these approaches in events of distinct natures.

In this work, we aim at building a unifying framework that is suitable for the classification of events of distinct natures. The framework has as key elements: a) external enrichment using related web pages for extending the conceptual features contained within the tweets; b) semantic enrichment using the Linked Open Data cloud to add related semantic features; and c) a pruning technique that selects the semantic features with discriminative potential.

We evaluated our proposed framework using a broad experimental setting, that includes: a) seven target events of different natures; b) different combinations of the conceptual features proposed (i.e. entities, vocabulary and their combination); c) distinct feature extraction strategies (i.e. from tweet text and web related documents); d) different methods for selecting the discriminative semantic features (i.e. pruning, feature selection, and their combination); and e) two classification algorithms. We also compared the proposed framework against another kind of contextual enrichment based on word embeddings.

The results showed the advantages of using the proposed framework, and that our solution is a feasible and generalizable method to support the classification of distinct event types.

Keywords: Semantic Web. DBPedia. LOD. Twitter. Event Classification.

Um Framework para Classificação de Eventos em Tweets Baseado em Enriquecimento Semântico Híbrido

RESUMO

As plataformas de Mídias Sociais se tornaram um meio essencial para a disponibilização de informações. Dentre elas, o Twitter tem se destacado, devido ao grande volume de mensagens que são compartilhadas todos os dias, principalmente mencionando eventos ao redor do mundo. Tais mensagens são uma importante fonte de informação e podem ser utilizadas em diversas aplicações. Contudo, a classificação de texto em *tweets* é uma tarefa não trivial. Além disso, não há um consenso quanto à quais tarefas devem ser executadas para Identificação e Classificação de Eventos em tweets, uma vez que as abordagens existentes trabalham com tipos específicos de eventos e determinadas suposições, que dificultam a reprodução e a comparação dessas abordagens em eventos de natureza distinta.

Neste trabalho, nós elaboramos um *framework* para a classificação de eventos de natureza distinta. O *framework* possui os seguintes elementos chave: a) enriquecimento externo a partir da exploração de páginas *web* relacionadas, como uma forma de complementar a extração de *features* conceituais do conteúdo dos tweets; b) enriquecimento semântico utilizando recursos da *Linked Open Data cloud* para acrescentar *features* semânticas relacionadas; e c) técnica de poda para selecionar as *features* semânticas mais discriminativas.

Nós avaliamos o *framework* proposto através de um vasto conjunto de experimentos, que incluem: a) sete eventos alvos de natureza distinta; b) diferentes combinações das *features* conceituais propostas (i.e. entidades, vocabulário, e a combinação de ambos); c) estratégias distintas para a extração de *features* (i.e. a partir do conteúdo dos tweets e das páginas *web*); d) diferentes métodos para a seleção das *features* semânticas mais relevantes de acordo com o domínio (i.e. poda, seleção de *features*, e a combinação de ambos); e) dois algoritmos de classificação. Nós também comparamos o desempenho do *framework* em relação a outro método utilização para o enriquecimento contextual, o qual tem como base *word embeddings*.

Os resultados mostraram as vantagens da utilização do *framework* proposto e que a nossa solução é factível e generalizável, dando suporte a classificação de diferentes tipos de eventos.

Palavras-chave: Web Semântica. DBpedia. LOD. Twitter. Classificação de Eventos.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BOW	Bag-of-Words
GloVe	Global Vectors
LDA	Latent Dirichlet Allocation
LOD	Linked Open Data
LSA	Latent Semantic Analysis
OWL	Web Ontology Language
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
POS	Part-of-Speech
RDF	Resource Description Framework
RDFS	RDF Vocabulary Definition Language
SMO	Sequential Minimal Optimization
SVM	Support Vector Machines
TF-IDF	Term Frequency – Inverse Document Frequency
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

CONTENTS

1 INTRODUCTION

Social media platforms have become key as a means of spreading information, opinions or awareness about real-world events (MCMINN; MOSHFEGHI; JOSE, 2013; MEDVET; BARTOLI, 2012). Among the most popular platforms, Twitter stands out for its large number of users who, together, produce more than 500 millions[1] of daily messages about all sorts of topics and subjects. Such messages are an important source of useful information about events of all types and magnitude (BECKER; NAAMAN; GRAVANO, 2011; SCHULZ; RISTOSKI, 2013; SAKAKI; OKAZAKI; MATSUO, 2010). In addition, they include the objective and/or subjective perspective of different users. The detection and classification of event-related tweets have many useful applications, such as the identification of breaking news, real-time awareness, and updates about events (e.g. car crashes, political protests, fires, natural disasters, epidemics), measurement of the repercussion of a given event either through the volume of messages, or the perception the population have towards them (e.g. sentiment) (SANKARANARAYANAN et al., 2009; SAIF; HE; ALANI, 2012; SAKAKI; OKAZAKI; MATSUO, 2010; PACKER et al., 2012; SCHULZ; RISTOSKI; PAULHEIM, 2013; ARAMAKI; MASKAWA; MORITA, 2011).

However, text classification on Twitter is by no means a trivial task that can be handled by conventional Natural Language Processing (NLP) techniques (BECKER; NAAMAN; GRAVANO, 2010; KHUC et al., 2012; CAMBRIA et al., 2013). Compared to the classification of longer, structured documents, tweet classification faces additional challenges. First, by design Twitter messages contain little textual information and several tricks are used to convey meaning (e.g. encoded URLs with additional information, hashtags, abbreviations, emoticons), resulting in very noisy text pieces. Furthermore, they often exhibit low quality (e.g. typos, ungrammatical sentences) and contain a very informal and dynamic vocabulary (MCMINN; MOSHFEGHI; JOSE, 2013; SAIF; HE; ALANI, 2012). Second, in addition to massive scale, Twitter users post messages with a variety of content types, which differ in subject, scope, and purpose. Finding events of interest in this diverse, sparse volume of messages is thus a challenging problem. Due to these characteristics, it can be necessary to apply filtering techniques to first separate and then organize the tweets according to the topic addressed. The challenge in this field is to identify whether the tweets clustered in the same group belong to the same event or were published in a specific range time (MCMINN; MOSHFEGHI; JOSE, 2013; BECKER et al., 2012; PACKER et al., 2012).

In this context, the Event Identification and Classification field emerged with the goal of

[1] https://about.twitter.com/company

identifying whether Twitter messages are associated with a specific event (e.g. the Hurricane Sandy) or domain (e.g. natural disasters) (ATEFEH; KHREICH, 2015; SAKAKI; OKAZAKI; MATSUO, 2010). Despite the lack of consensus, most works address *event identification* as the overall task of creating groups of subject and time-related tweets from large, never-ending data streams (SAKAKI; OKAZAKI; MATSUO, 2010; PACKER et al., 2012), within which the *event classification* is a specific task that deals with the construction of classification models that filter and categorize events (SCHULZ; RISTOSKI; PAULHEIM, 2013; REUTER; CIMIANO, 2012; SCHULZ; GUCKELSBERGER; JANSSEN, 2015). The present work assumes such a distinction and focuses on event classification.

Existing approaches for event classification often focus on specific types of events, such as epidemics (TSOU et al., 2015; ARAMAKI; MASKAWA; MORITA, 2011), incidents (ABEL et al., 2012b; SCHULZ; RISTOSKI, 2013), and natural disasters (SAKAKI; OKAZAKI; MATSUO, 2010). They also rely on assumptions involving the volume of posts (LI; SUN; DATTA, 2012), temporal and geo-spatial properties (e.g. small-scale incidents and crisis situations) (SCHULZ; RISTOSKI; PAULHEIM, 2013; ANANTHARAM et al., 2015), agents involved, or the vocabulary used (PACKER et al., 2012; MEDVET; BARTOLI, 2012; BECKER et al., 2012). This wide variety of assumptions makes it difficult to reproduce and compare these approaches in events of distinct natures.

As a means to deal with the poor textual content of tweets, related work has suggested the use of external information to add context to tweet contents, in applications such as Event Classification (ABEL et al., 2012a; PACKER et al., 2012; SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015) and Sentiment Analysis (SAIF; HE; ALANI, 2012). As a general approach, the textual features to be enriched are selected according to some criterion, and mapped into the resources available in external knowledge source (e.g. Wikipedia, Linked Open Data cloud). Then, related contextual properties (e.g. semantics, co-occurrences, representative terms) are retrieved and combined with the textual features. Criteria used to select textual tokens to be enriched are named entities (SAIF; HE; ALANI, 2012; ABEL et al., 2012a; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; VOSECKY et al., 2014), frequent or relevant terms measured using *Term Frequency – Inverse Document Frequency* (TF-IDF) (PACKER et al., 2012; SCHULZ; RISTOSKI; PAULHEIM, 2013; VOSECKY et al., 2014), and location/time identification heuristics (SCHULZ; GUCKELSBERGER; JANSSEN, 2015; VOSECKY et al., 2014). For the contextual enrichment, different techniques and elements can be explored:

- External documents: context is provided by the content extracted from related web docu-

ments, which can be identified through Uniform Resource Locators (URL) mentioned in the tweets (VOSECKY et al., 2014), or by selecting specific words in the messages (e.g. named entities, representative terms) to access, for example, related Wikipedia pages (GENC; SAKAMOTO; NICKERSON, 2011; ROSA et al., 2011). The challenges of this approach are defining which content to extract, and which data is important for the event classification process;

- Semantic web: context is provided by mapping elements extracted from the tweets into resources available in the Linked Open Data (LOD) cloud, which contains semantic properties describing resources from different domains (e.g. user-generated content, media, cross-domain). Semantic properties can help generalizing the contents of the tweets (ABEL et al., 2012a; PACKER et al., 2012; SCHULZ; GUCKELSBERGER; JANSSEN, 2015), or expanding to related concepts. However, there are no guidelines to determine which textual features to enrich, nor which knowledge sources (e.g. DBpedia, YAGO) and semantic properties to adopt (e.g. *rdf:type*, *dct: subject*);

- NER tools: indirect sources connected to Named Entity Recognition (NER) tools (e.g. Open Calais, Alchemy, DBpedia Spotlight) are employed to identify entities that are mentioned in the tweet text and provide related categorical information (SAIF; HE; ALANI, 2012; VOSECKY et al., 2014). Some NER tools are related to proprietary knowledge sources (e.g. Open Calais), whilst others also explore open knowledge sources (e.g. Alchemy). The main challenge of this approach is selecting the relevant content provided by these tools without distorting the focus of the event analyzed by the inclusion of unrelated information.

Nevertheless, each work assumes a particular definition of event for undertaking the contextual enrichment, underlined by specific assumptions, and a particular application purpose, which is not necessarily generalizable. Moreover, the aforementioned contextual enrichment techniques can be combined with variations on the textual features, knowledge bases, properties, and NER tools. Therefore, it is difficult to reproduce, compare and select among the different enrichment approaches. Also, semantic enrichment results in a huge amount of new features, most of which have no discriminative power for event classification (JANPUANGTONG; SHELL, 2015; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; ROMERO; BECKER, 2016a).

In this work, we leverage these previous enrichment experiences on selecting relevant textual features and enrichment strategies using external sources to build a unifying framework

for the classification of event-related tweets, such that the classification of events of distinct types can be performed and compared. The focus of this work is the contribution of semantic enrichment, possibly in combination with other enrichment strategies.

The goals of this work are:

- to identify distinct event definitions and the features used to characterize them;
- to identify external information resources that can be used to enrich tweet contents with contextual information, according to the different event types;
- to define a process to semantically enrich the contents of event-related tweets, as a means to improve the event classification, which is applicable to events of different types;
- to develop experiments to measure the contribution of semantic enrichment for event-related tweet classification, possibly in combination with other enrichment strategies.

According to the goals aforementioned, the research questions explored in this work are described below:

- are there specific features that, if enriched, are more discriminatory of certain types of events?
- what is the enrichment process that yields the best results for the event classification in tweets and how to apply it?

We aim at building a unifying framework that is suitable for the classification of events of distinct natures, ranging from planned (e.g. concerts, sports) to unplanned (e.g. incidents, natural disasters) events. We propose a hybrid enrichment process in which we combine semantic enrichment with two other contextual enrichment strategies, namely external source and NER. These strategies are complementary, as follows:

- NER: help recognizing in tweets and external documents, the entities that are relevant for event characterization, namely the agents and locations involved in the events;
- External source enrichment: to improve the identification of more representative terms or entities related to the event, thus helping to overcome the poor and sparse textual contents of tweets;
- Semantic enrichment: by exploring the semantic properties available in the LOD cloud, we generalize all this information and obtain more domain representative concepts, to help in the event classification problem.

The framework also handles a second related problem, which is how to select from this huge volume of semantic features the most discriminative ones concerning the type of event at hand. We propose a pruning method based on the PageRank algorithm (PAGE et al., 1999), to be used in combination with other feature selection methods (LIU; YU, 2005), as a means to select more discriminative semantic features and improve the classification of events.

We evaluated the proposed framework using a broad experimental setting that includes:

- several datasets that represent events of distinct natures, ranging from sportive events to natural disasters, and epidemics;
- the combination of different types of textual features extracted from tweet texts;
- the contribution of the semantic enrichment, and the benefits of combining it with external source enrichment;
- the contribution of a generic feature selection technique (i.e. the correlation-based feature selection algorithm) (LIU; YU, 2005), and specific-purpose semantic feature pruning technique, the latter with different pruning thresholds;
- two different classification algorithms widely used in text classification problems, namely Naïve Bayes (NB) and an implementation of Support Vector Machines (SVM) called Sequential Minimal Optimization (SMO) (RUSSELL; NORVIG, 1995; PLATT, 1998).

We performed the experimental evaluation using two different baselines, the first one based on tweet textual features only and the second one representing another alternative of contextual enrichment, namely word embeddings (KENTER; RIJKE, 2015; LI et al., 2016).

Analyzing the performance of our proposed framework, considering the different setups combinations, we were able to statistically outperform the baseline in 25.4% of cases. In general, improvements could be noticed in 53.17% of the results, with a maximum improvement of 32.6 percentage points. The datasets composed of named entities were the ones that presented the best results. Regarding the word embeddings comparison, the results showed that our framework was able to statistically outperforms this new baseline in 83% of cases, ensuring the efficiency of our approach.

In summary, the results showed the advantages of using the proposed hybrid semantic enrichment framework, and that our solution is a feasible and generalizable method to support the classification of distinct event types.

Our contributions can be summarized as follows:

- we propose a set of conceptual features to be semantically enriched, namely agents, location, frequent and representative terms (ROMERO; BECKER, 2016a);
- we propose a hybrid semantic enrichment process that extracts these features from both tweet texts and related web documents (e.g. news sites mentioned in the tweets), and map them into resources described in the LOD cloud to retrieve semantic properties;
- we propose a pruning method to select potentially relevant and discriminative semantic features, based on the adaptation of the PageRank algorithm;
- we develop experiments using several datasets reporting events of distinct natures, in which we compare the contributions of semantically enriched features, the use of external content in addition to tweet contents, as well as the proposed pruning method for the selection of discriminative features;
- we compare our proposed framework with an alternative contextual enrichment approach, based on word embeddings.

The rest of this work is structured as follows. In Chapter 2 the background needed for the understanding of the work is described. In Chapter 3 we present an overview of related work. Chapter 4 highlights the main aspects of the proposed approach. Chapter 5 provides a detailed description of the experiments performed and their results. Conclusion and future work are addressed in Chapter 6.

2 BACKGROUND

This chapter presents the background needed for understanding the proposed work. We present the definitions and concepts related to the term event and the features that characterize it; the concepts, technologies, and techniques related to Semantic Web; the classification algorithms, and metrics that will be used to classify events and analyze the results according to the approach proposed; as well as the word embeddings approach.

2.1 Event Definition and Categorization

Works on Event Identification and Classification in tweets adopted definitions that highlight different aspects of an event. These definitions usually rely on the topic, purpose, and the scope of the event classification task that will be executed. Such diversity has originated distinct methods and techniques, making it difficult to compare and extend these different approaches to other event types.

Atefeh and Khreich (2015) classify events as *specified* or *unspecified*, where the former benefits from the existence of prior information (e.g. description), and the latter relies on monitoring trends or the sudden burst of a group of expressions, most often at real-time (MCMINN; MOSHFEGHI; JOSE, 2013; MEDVET; BARTOLI, 2012; SANKARANARAYANAN et al., 2009; LIU et al., 2016; NOURBAKHSH et al., 2015).

Specified and unspecified events can be additionally classified as *planned* and *unplanned* (BECKER et al., 2012; SAKAKI; OKAZAKI; MATSUO, 2010; SCHULZ; RISTOSKI; PAULHEIM, 2013; PACKER et al., 2012; SANKARANARAYANAN et al., 2009; LIU et al., 2016). Similarly to specified events, the classification of planned events rely on prior event-related information (BECKER et al., 2012; PACKER et al., 2012), whereas unplanned events are related to incidents and natural disasters (SAKAKI; OKAZAKI; MATSUO, 2010; SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015).

In addition to these two orthogonal categories for events, others features can help in the Event Identification and Classification task. Table 2.1 presents a non-exhaustive list of event definitions proposed by the related work. All definitions are directly or indirectly related to a *topic* or *subject*, which is either the content to be discovered as part of the Event Identification and Classification task (Defs. 1, 2, and 4), or an input to guide the filtering, processing and the event classification of specified (types of) events (Defs. 3, 5). The subject of an event can be characterized by factors such as similarity of terms (Defs. 1, 2), specific vocabulary (Defs. 3,

Table 2.1: Definitions of event

Id	Definition	Purpose	Application	Main Concepts
1	A real-world occurrence *e* with (1) an associated time period *Te* and (2) a time-ordered stream of Twitter messages *Me*, of substantial volume, discussing the occurrence and published during time *Te* (BECKER; NAAMAN; GRAVANO, 2011)	On-line identification of real world event content	General (unspecified) event identification and classification	Topic, time, and scale
2	An event is a significant thing that happens at some specific time and place. Something is significant if it may be discussed in the media (MCMINN; MOSHFEGHI; JOSE, 2013)	Methodology for the automatic creation of an event detection corpus	General (unspecified) event identification and classification	Topic, time, and location
3	Events have several properties: i) they are of large scale (many users experience the event), ii) they particularly influence peoples daily life (for that reason, they are induced to tweet about it), and iii) they have both spatial and temporal regions (so that real-time location estimation is possible). An event might have actively participating agents, passive factors, products, and a location in space/time (SAKAKI; OKAZAKI; MATSUO, 2010)	Real-time monitoring of disastrous events (e.g. earthquakes, storms, fires, traffic jams)	Natural disaster events monitoring (specified)	Topic, time, scale, and agents
4	An event is a theme of conversation that becomes suddenly popular amongst tweets of the same topic (MEDVET; BARTOLI, 2012)	Detecting and summarizing popular events related to a given general topic (brand)	General (unspecified) event identification and classification	Topic and scale
5	5-tuple (^etype, ^eloc, ^est, ^eet, ^eimpact), where ^etype, refers to the event type such as accident, breakdown, and music event; ^eloc, refers to the location of the event (lat-long); ^est, and ^eet, refer to the start time and end time of the event; and ^eimpact, refers to a number quantifying the severity of the event (ANANTHARAM et al., 2015)	Leverage citizen observations as a source of city events	General (specified and unspecified) event identification and classification	Topic, time, scale, and location

Source: the author.

5), burstiness of vocabulary (Def. 4), among others.

The *temporal component* is present in almost all definitions (Defs 1, 2, 3, and 5), determining the period of the event occurrence. *Scale* is another property often highlighted (Defs. 1, 3, 4, and 5), with the assumption that a high volume of messages conveys the relevance of the event that motivates people to report it (large-scale events). However, there are works that specifically aim to identify local events (e.g. car crashes, fire), of which the impact is small in terms of an absolute number of posts, but significant considering a small community (SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015).

Geographical properties are highlighted in definitions that involve incident management (Def. 3), which not necessarily can be anticipated (e.g. earthquakes, epidemics, and car crashes), and are often referred to as emergency or unplanned events. The geographical property also appears in the Defs. 2 and 5, which deals with events of different domains, and consider

the location where the event took place as an important feature to be analyzed. In addition, Def. 6 exposes that city events may be planned (e.g. cultural event, service maintenance) or unplanned (e.g. traffic related events). Def. 3 also includes passive or active *agents* (e.g. people, organizations, or geopolitical areas) that are involved or affected by the event.

Based on the definitions presented in Table 2.1, in this work we consider an event as:

Definition 2.1. *An event is an occurrence, represented by a topic, that occurs in a specific time and can involve one or more locations and agents.*

According to *Definition 2.1*, we are able to cover several types of event, such as sports, commemorative dates, epidemics, natural disasters, and incidents in different scales, which can attend the focus of our work that is specified events. Regarding the features that characterize an event, the different definitions described in Table 2.1 present distinct concepts that can be the object of enrichment. We use these definitions to select a set of core features to be extracted and enriched, as described in Chapter 3.

2.2 Semantic Web and the Linked Open Data Project

The Semantic Web was designed for connecting data through semantic relations, enabling the utilization of these data for humans and machines (ABELLÓ et al., 2015). Instead of just displaying data available on the Web, the Semantic Web aims at enabling machines to analyze, infer the relationships between the facts, and comprehend the data content available on the Web (BERNERS-LEE; HENDLER; ORA, 2001).

Linked Data refers to a set of best practices to structure and add semantics to the traditionally way used to represent the data published on the web (BIZER; HEATH; BERNERS-LEE, 2009). By converting the data to a machine-readable format, these best practices can enable the elaboration of new types of applications, through the connection of data from different domains. The Linked Data principles are:

- use Uniform Resource Identifier (URI) as name for things;
- use HTTP URIs to look up those names;
- provide useful information by using standards (e.g. RDF, SPARQL);
- include links to other URIs to discover more things.

A URI is a name for things on the Web and enables the identification of any entity in the world. The Resource Description Framework (RDF) is a format that provides a graph-based data model to structure and link things in the world. The RDF format encodes data as triples composed of *subject*, *predicate*, and *object*. The predicate refers to how the subject and the object are related. Both the subject and the object represent an entity that is identified by a URI (BIZER; HEATH; BERNERS-LEE, 2009).

To describe entities and how they are related, different vocabularies can be used. Vocabularies are collections of classes and properties that are used to describe and model different domains of interest. The RDF Vocabulary Definition Language (RDFS) and the Web Ontology Language (OWL) provide a basis for creating specific vocabularies (BIZER; HEATH; BERNERS-LEE, 2009).

Using standard vocabularies, one is able to represent formal natural language expressions in a machine-readable way. For example, we can use the RDFS vocabulary to represent classes and resources (i.e. *rdfs:Class* and *rdfs:Resource*). Given that, we are able to translate the expression *"Mary Poppins is a person"* into an instance of the triple *rdfs:Resource rdf:type foaf:Person*. In this example, *Mary Poppins* is a resource available in the knowledge base (i.e. instance of *rdfs:Resource*), *is* represents the predicate using the property *rdf:type*, and *person* is a class in the Friend of a Friend (FOAF) vocabulary (*foaf:Person*).

2.2.1 Linked Open Data Project

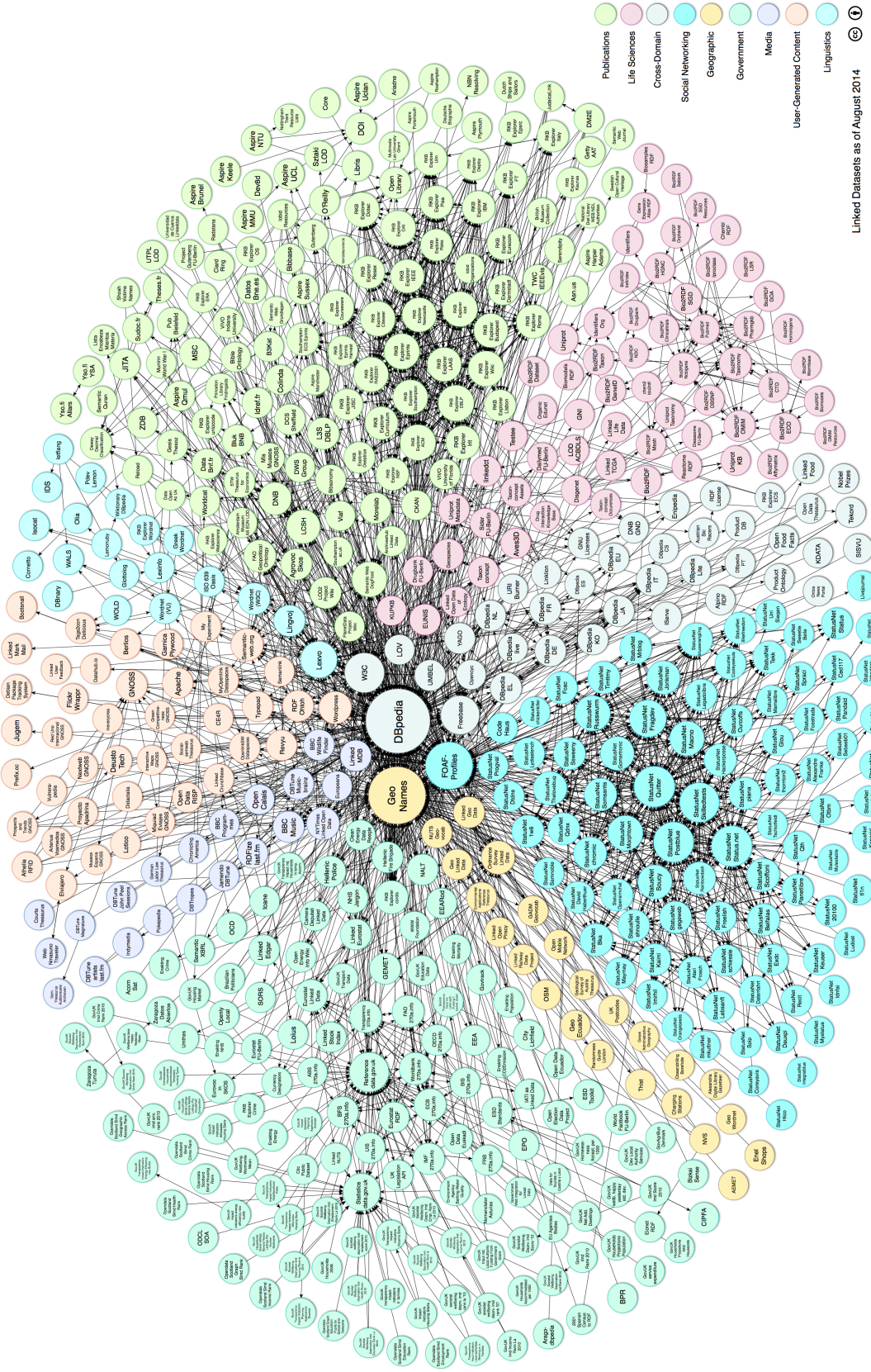
The LOD Project was founded in 2007 as an effort to publish datasets in RDF format and under open licenses, according to the Linked Data principles and interlinking it with existing datasets (BIZER; HEATH; BERNERS-LEE, 2009; SCHMACHTENBERG; BIZER; PAULHEIM, 2014).

Figure 2.1 represents the LOD cloud diagram as mapped in August of 2014. Each node on the cloud diagram represents a different dataset, and the arcs indicate the existence of links between the resources of distinct datasets. The size of the circles reflects the indegree (i.e. number of datasets that point to a specific dataset) of the corresponding dataset.

The LOD cloud is composed of data from different domains. According to Schmachtenberg, Bizer and Paulheim (2014), these domains can be classified into topical categories, as follows:

- Media: provides information about films, music, TV and radio programs;

Figure 2.1: Linked Open Data cloud



Source: http://lod-cloud.net/.

- Government: statistical data provided by federal and local governments;
- Publications: provides information about scientific publications and conferences;
- Life sciences: provides biological and biochemical information, drug-related data, and information about species and their habitats;
- Geographic: provides information about geographic entities and geopolitical divisions;
- Cross-domain: provides information of general knowledge;
- User-generated content: provides data from portals that collect content generated by larger user communities;
- Social networking: provides data that describes people profiles and describes the social ties between people;
- Linguistic: provides open linguistic data.

As we can observe in Figure 2.1, there are datasets that serve as hubs, providing general knowledge, while others act as a kind of authority, providing information about specific domains. For example, the DBpedia knowledge base consists of RDF triples extracted from infoboxes of Wikipedia articles, and Geonames knowledge base provides RDF descriptions of millions of geographical locations in the world (AUER et al., 2007; BIZER et al., 2009; SCHMACHTENBERG; BIZER; PAULHEIM, 2014).

In this work, we focus on the cross-domain topical category, through which we will add generalizations to all the textual concepts used to describe event-related tweets.

2.2.2 DBpedia

The DBpedia project focuses on extracting structured Wikipedia content and provides access to this information on the Web (AUER et al., 2007; BIZER et al., 2009). According to information provided by its wiki[1], the DBpedia knowledge base currently describes over 4.58 million instances in 125 languages, including 1,445,00 people, 735,000 places, 411,000 creative works (i.e. music albums, films, and video games), 241,000 organizations, 251,000 species and 6,000 diseases.

[1] http://wiki.dbpedia.org/about

The content of Wikipedia articles is composed mostly of free text and images. Structured information can be found in infobox templates (i.e. a table in the format of attribute-value pairs), such as categorization information, geo-coordinates, links to external Web pages, and links to the different language editions of Wikipedia (AUER et al., 2007; BIZER et al., 2009).

For each resource available in the DBpedia knowledge base, a URI reference in the format of *http://dbpedia.org/resource/Name* is defined, where Name is derived from the URL of the source Wikipedia article. This identifier can dereference the resource into a rich RDF description, which includes human-readable definitions, relationships to other resources, classification in the concept hierarchies, and links to other sources that also describe this resource (AUER et al., 2007; BIZER et al., 2009). For example, a Wikipedia article for the soccer player Didier Drogba can be found on the link *https://en.wikipedia.org/wiki/Didier_Drogba*. The correspondent information based on the Linked Data principles can be found on the URI *http://dbpedia.org/page/Didier_Drogba*.

In this work, we selected the DBpedia knowledge base as the source for semantic enrichment, since it is a cross-domain knowledge base that covers a huge amount of information. In addition, it contains in/out links to an expressive number of other datasets on the LOD cloud.

2.2.3 SPARQL

The SPARQL Protocol and RDF Query Language (SPARQL) is a semantic query language that enables the recovery and manipulation of data in the RDF format (TECHENTIN et al., 2014). As aforementioned, the RDF format consists of triples containing a subject, a predicate, and an object. SPARQL can be used to elaborate queries to retrieve RDF graphs (i.e. a set of RDF triples) across diverse sources (TECHENTIN et al., 2014).

An example of an SPARQL query is presented in Figure 2.2. This query retrieves the object of all triples in which the DBpedia resource Didier Drogba is the subject and *rdf:type* is the predicate. SPARQL keywords are in uppercase.

In this work, we employ SPARQL queries to retrieve semantic properties related to textual features extracted from the tweets text and related web documents. The queries are elaborated as required, for example, to identify the resources to perform the semantic enrichment step and to create a network of semantic features, based on the relationship among the objects recovered for each resource. We use the SPARQL endpoint *dbpedia.org/sparql*.

Figure 2.2: SPARQL query example

```
PREFIX dbres: < http://dbpedia.org/resource/>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX   rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?type WHERE {
    dbres:Didier_Drogba rdf:type ?type.
  }
```

Source: the author.

2.3 Text Classification and Metrics

Text classification is the problem related to the organization and categorization of textual data. This problem relies on NLP techniques and can be employed in different types and formats of text. The levels of textual analysis can also differ according to the application purpose (CAMBRIA et al., 2013). In this work, we consider two different types of text: tweets and web documents (i.e. news sites, blogs, and also other tweets). In both cases, the analysis is performed at document level.

To execute text classification, NLP techniques are employed to identify and extract features that better describe the event being analyzed. If necessary, feature selection and pruning methods can also be adopted to determine which of these features are more relevant for the analysis. After that, Machine Learning algorithms are trained to classify the texts. According to the application purpose, different evaluation metrics are employed to compare the results.

2.3.1 NLP and Information Retrieval Techniques

The content of tweet text is very informal and several tricks are used by Twitter users to convey meaning to their posts, such as hashtags, emoticons, and abbreviation. To identify the important pieces that compose a Twitter message, different approaches can be employed (SAIF; HE; ALANI, 2012; MOHAMMAD; KIRITCHENKO; ZHU, 2013; SCHULZ; RISTOSKI; PAULHEIM, 2013; ABEL et al., 2012a; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; VOSECKY et al., 2014; PACKER et al., 2012), among them:

- tokenization to split the sentence according to a specific separator, revealing the terms and symbols that compose the tweet;
- normalization to convert similar terms and symbols into a specific representation (e.g. to convert @user, URLs, and emoticons into symbols such as T_USER, T_URL, and

T_EMOT);

- term weighting techniques (e.g. TF-IDF) to identify the relevance of each term in the event-related Twitter dataset;
- Stemming to obtain the inflected word, through the removal of affixes and other letters used to define gender and to differentiate the verbal tense of a word, for instance;
- Part-of-Speech Tagging (POS) to categorize words in the message (e.g. nouns, verbs, adjective). According to the application purpose, only words belonging to specific categories are employed (e.g. nouns);
- NER techniques to identify named entities that appear in the text and classify them in a specific category, such as person, organization, location, or monetary values. Barack Obama/Person, IBM/Organization, and Brazil/Location are examples of named entities that can be recovered through this technique. Popular NER tools are Open Calais[2], Alchemy API Entity Extraction[3], and Zemanta[4].

In this work, we adopted named entities, frequent and representative terms as the textual features to be extracted from documents, and used as input for the proposed hybrid semantic enrichment process.

2.3.2 Algorithms for Text Classification

To classify a Twitter message according to its topic, different approaches can be applied. Supervised learning algorithms, such as Naïve Bayes, SVM, Random Forest, JRip, and Maximum Entropy are very popular (SCHULZ; RISTOSKI; PAULHEIM, 2013; SAIF; HE; ALANI, 2012; MOHAMMAD; KIRITCHENKO; ZHU, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; ARAMAKI; MASKAWA; MORITA, 2011; SAKAKI; OKAZAKI; MATSUO, 2010). These algorithms are mostly used to classify the tweet as belonging or not to a specific event (e.g. an earthquake, an incident, a tweet about influenza).

Clustering techniques and Neural Networks are also applied, mainly for applications in which it is necessary to organize the tweets according to its characteristics and similarities, such as organizing breaking news according to its category (e.g. sports, politics) (FISICHELLA et al., 2011; ROWE; STANKOVIC, 2011; LIU et al., 2016).

[2] http://www.opencalais.com/
[3] http://www.alchemyapi.com/
[4] http://www.zemanta.com/

In this work, we employed the supervised algorithms NB and an implementation of SVM called Sequential Minimal Optimization (SMO). The NB classifier belongs to the family of the probabilistic classifiers. For the classification, it considers a vector of features values and assumes that these features are independent (RUSSELL; NORVIG, 1995). The SVM classifier is a non-probabilistic linear classifier that represents the features in a high dimensional space, and then constructs hyperplanes to better separate and classify the data. The SMO is an optimized implementation of SVM, in which the problem is broken down into small pieces to be solved analytically (PLATT, 1998).

2.3.3 Feature Selection Algorithms

Feature Selection algorithms have the goal of selecting the most relevant features in a given dataset, according to specific criteria. In general, it aims at removing redundant and irrelevant features from the dataset. By employing this kind of technique, we are able to better represent the dataset, which will be composed of the most discriminative features. In addition, with a small set of features, we can improve the machine learning algorithms performance, as well as reduce the computational cost of the training step.

There are several feature selection algorithms (LIU; YU, 2005). In this work we employed two of them: a) an algorithm based on information gain, which ranks the features according to the information gain produced (i.e. entropy reduction), and b) the Correlation-based feature selection (CFS) algorithm, in which a set of features is considered good if it contains feature highly correlated with the class and not correlated with each other.

The use of this type of technique aims at reducing the number of features by selecting the most discriminative ones, so as to improve the classification performance.

2.3.4 Evaluation Metrics

Table 2.2: Confusion Matrix

		Predicted Condition	
		C_1	C_2
True Condition	C_1	True Positive (TP) C_1	False Negative (FN) C_2
	C_2	False Positive (FP) C_1	True Negative (TN) C_2

Source: the author.

To analyze the results of a classifier, the values that compose the Confusion Matrix are considered. These values represent the amount of correct and incorrect instances classified by the algorithms. Table 2.2 shows the True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) values that compose the Confusion Matrix.

Based on the results from the Confusion Matrix, it is possible to calculate the measures *Precision*, *Recall*, and *F-Measure*, for each class, which represent the performance of the classifier. The Precision metric represents the amount of instances that were correctly classified as belonging to a given class by the algorithm, as presented in Equation 2.1. Given a specific class, the Recall metric represents the correctly classified instances for that class, as described in Equation 2.2. The F-Measure metric represents the harmonic mean between Precision and Recall, in which the same relevance is considered for both metrics, as described in Equation 2.3. We can also calculate the weighted measure for the set of classes. For example, the weighted F-Measure is presented in Equation 2.4. The weighted version of the equation can also be calculated to Precision and Recall.

In addition to these metrics, which allow us to compare the performance of each classifier, we also evaluated the statistical significance of the differences between our approach and the approaches used as the baseline. For that purpose, we employed a statistical Student's t-test (CALLEGARI-JACQUES, 2009), adopting a significance level of 0.05.

$$Precision_{Ci} = \frac{TP_{Ci}}{TP_{Ci} + FP_{Ci}} \tag{2.1}$$

$$Recall_{Ci} = \frac{TP_{Ci}}{TP_{Ci} + FN_{Ci}} \tag{2.2}$$

$$F - Measure_{Ci} = 2 \times \frac{Precision_{Ci} \times Recall_{Ci}}{Precision_{Ci} + Recall_{Ci}} \tag{2.3}$$

$$F\text{-}Measure = \frac{\#c_1 \times F\text{-}Measure_{c_1} + \#c_2 \times F\text{-}Measure_{c_2}}{\#c_1 + \#c_2} \tag{2.4}$$

where $F\text{-}measure_{c_i}$ is the F-Measure for class i, and $\#c_i$ is the number of elements in the class.

2.4 Word Embeddings

Language representation though semantic vector space models have been explored as another approach that can be used in NLP applications, such as information retrieval, text clas-

sification and question answering (KENTER; RIJKE, 2015; LI et al., 2016). This distributional semantic approach is called word embeddings, due to the word vectors produced for each word in the vocabulary. Given a set of textual data, it represents each word in a dimensional space, in which the proximity to other similar words is given by mathematical properties that are able to connect these features according to the context (i.e. the closest words in a dimensional space are semantically or syntactically similar). Thus, this word-based model is able to capture the semantic similarity between words, producing a sense of meaning, rather than considering only words that have similar letters (KENTER; RIJKE, 2015; PENNINGTON; SOCHER; MANNING, 2014; MIKOLOV et al., 2013).

Different strategies can be used to build the vector space and produce a model to be used in the NLP tasks. Mikolov et al. (2013) proposed the word2vec algorithm, which is based on two layers neural networks which after training produces a vectorial representation of words which conserve the linguistic contexts of each word. Two architectures can be used to produce the model, the continuous bag-of-words (CBOW), in which a term is predicted according to its surrounding terms, and Skip-gram, in which one term is used to predict multiple terms. Another strategy is Global Vectors (GloVe), which is based on a global matrix factorization where the terms co-occurrences are considered (i.e. it consider how frequently words co-occur with one another in a given corpus) (PENNINGTON; SOCHER; MANNING, 2014). To build the model, a huge corpora is needed, such as the Google News dataset, the UMBC web base corpus, a Wikipedia dump, a huge volume of tweets, among others (LI et al., 2016).

Given the widely use of this approach, the semantic characteristics, and the good results reported for topic modeling in event identification, contextual word similarity and text classification (KENTER; RIJKE, 2015; LI et al., 2016; LIU et al., 2015), we decided to use this approach as baseline in our analysis, which will be used in comparison to our proposed semantic enrichment approach.

2.5 Final Remarks

In this chapter, we presented the main definitions and concepts about Event Identification and Classification in Tweets, as well as the techniques, algorithms, and metrics that were used to developed and experiment the approach proposed in this work.

3 RELATED WORK

In this chapter, we examine the related work on the Event Identification and Classification field. These works are analyzed according to the contextual enrichment approach employed, the features used as input to the enrichment approach, and the algorithms used to perform the classification task. Finally, we summarize all approaches described and compare them to the one proposed in this work.

3.1 Motivation to Event Identification and Classification in tweets

The Event Identification and Classification field has emerged with the goal of identifying and organizing event-related documents, and therefore, it is also referred to *event detection* (MCMINN; MOSHFEGHI; JOSE, 2013). Researches in this field attempt to perform an event-based organization on stories, breaking news, and facts that happen around the world. Similar to clustering approaches, these efforts are aimed at categorizing the documents according to a predefined list of topics (e.g. sports, government, politics) or organizing the documents into groups according to the similarity among the subjects being addressed (MCMINN; MOSHFEGHI; JOSE, 2013; BECKER; NAAMAN; GRAVANO, 2011; ATEFEH; KHREICH, 2015).

Several efforts were directed at event detection on long documents, which are either structured or follow a formal language. An application example can be found in (FISICHELLA et al., 2011), which describes an approach to detect public health events on medical articles, by searching for specific elements on the document content: *who* (victims) was infected by *what* (diseases), *where* (locations) and *when* (time, defined as the period between the first relevant article and the last relevant one).

Regarding the detection and classification of events at real-time, the researchers attention have been directed to Social Media platforms, such as Twitter and Facebook, which have become key as a means of spreading information, opinions or awareness about real-world events. Specifically for Twitter, the object of study of this work, the main reasons for its wide use are (SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; SAKAKI; OKAZAKI; MATSUO, 2010; PACKER et al., 2012; BECKER; NAAMAN; GRAVANO, 2011):

- a huge amount of messages posted every day, which report a variety of events at real-time;
- tweets can report events of different types and scale, ranging from widely known events

(e.g. Olympics, US elections) to small-scale and local events (e.g. city traffic, car-crashes);

- Twitter messages reflect the point of view of the users who are interested in, participating at or witnessing the event occurrence;
- Twitter users can often report the occurrence of an event in anticipation to the traditional news media.

However, performing Event Identification and Classification in Twitter messages presents several challenges compared to event detection on longer structured documents (MCMINN; MOSHFEGHI; JOSE, 2013; BECKER; NAAMAN; GRAVANO, 2011), among them:

- twitter users post about 500 million of tweets every day, producing a huge volume of messages to be analyzed;
- most of the tweets report mundane and everyday issues, being necessary to apply filtering techniques to identify the messages that are relevant to the application purpose (i.e. the messages that report a real-world event);
- twitter messages contain little textual information, and several tricks are used to convey meaning (e.g. encoded URLs with additional information, hashtags, abbreviations, emoticons), resulting in very noisy text pieces.

Besides the challenges presented by the nature of tweets, the approaches proposed by the related work also display peculiarities that make it difficult to reproduce them in other types of events and compare the results:

- different features are used to characterize a real-world event (e.g. topic, agents involved, location);
- most of the works address specific event types, such as earthquakes (SAKAKI; OKAZAKI; MATSUO, 2010), epidemics (ARAMAKI; MASKAWA; MORITA, 2011), or are just interested in to identify breaking news on the Twitter posts (LIU et al., 2016);
- distinct techniques are employed to overcome the challenges presented by the Twitter messages (e.g. information extracted from Wikipedia articles, semantics extracted from LOD cloud datasets), but which are based on the adopted event definition and underlying assumptions.

3.2 Event Identification and Classification in Tweets Using Contextual Enrichment

As an attempt to overcome the challenges related to the poor textual and sparse nature of the messages posted by Twitter users, related work has suggested employing contextual enrichment, by using external information (ABEL et al., 2012a; PACKER et al., 2012; SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015).

One approach is using content extracted from related web documents, which can be identified through URLs mentioned in the tweets, or by selecting specific words in the posts to access, for example, related Wikipedia pages (GENC; SAKAMOTO; NICKERSON, 2011; ROSA et al., 2011; VOSECKY et al., 2014; SANKARANARAYANAN et al., 2009). In this work, we refer to this as *external documents enrichment*. The challenges of this approach are defining which content to extract and which data is important for the event classification process.

Another approach is contextual enrichment using information extracted from the Semantic Web, such as the knowledge bases provided by the LOD cloud (SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; ROWE; STANKOVIC, 2011), referred to as *semantic enrichment*. However, there are no guidelines to determine which textual features to enrich, nor which knowledge bases and properties to adopt, as each work assumes a particular definition of event that is related to the purpose of the application.

A third and more primitive approach is to employ NER tools to identify named entities that are mentioned in the tweet text and external documents, as a way to complement the enrichment through the use of specific categories of terms, such as person, organization, etc (ABEL et al., 2012b; VOSECKY et al., 2013; ROWE; STANKOVIC, 2011).

Related work also presented approaches that combine at least two of these techniques (i.e. external source, semantics, and NER tools) (PACKER et al., 2012; ABEL et al., 2012b; LI et al., 2016; SCHULZ; GUCKELSBERGER; JANSSEN, 2015), which we denominate as *hybrid enrichment*. In this section, we focus on external documents and semantic enrichment, since the NER tools strategy is commonly used to complement the enrichment, for example, through the recognition of named entities in tweet texts or web documents, which will be used as input to retrieve associated knowledge from the LOD cloud. More information about these techniques is presented below.

3.2.1 External Documents Enrichment

In related work, external web documents (e.g. news sites, sportive sites, blogs) have been used to help to identify breaking news (SANKARANARAYANAN et al., 2009), as well as to detect and classify general purpose events (VOSECKY et al., 2014; ROSA et al., 2011). Wikipedia articles have also been employed to perform tweet categorization (GENC; SAKAMOTO; NICKERSON, 2011).

To improve the performance and alleviate the inherent feature sparsity of tweets, the approach described by Rosa et al. (2011) proposes to access the pages mentioned in each post and retrieve its content to incorporate into the tweets. However, neither specific textual features were defined to be recovered, nor feature selection techniques were applied, which caused a decrease in the performance of the classifiers, specifically for precision and recall, since all content from the web documents was embedded into the messages.

Also focusing on general purpose events, the research developed by Vosecky et al. (2014) assumes that bag-of-words (BOW) techniques are not enough to retrieve the semantic and temporal aspects present in tweets contents. Thus, the proposed approach combines social and auxiliary semantics, by exploring five elements: the named entities person, organization, and location; general terms; and the trend behavior of the topic (i.e. timestamps). Social semantics is based solely on the textual content of the tweets, which is identified by analyzing the co-occurrence of specific hashtags and terms. The auxiliary semantics consist of mapping all URLs mentioned in the posts to extract named entities and top-k terms to be incorporated to the tweets.

TwitterStand is a system for the identification and clustering of breaking news on Twitter (SANKARANARAYANAN et al., 2009). It considers that Twitter users can use some artifacts, such as a link to other web pages, images, and video to complement the information about the breaking news. The aim of the system is thus not related to being the first one to detect the breaking news, but rather to recover tweets related to it. To define whether a tweet is an important breaking news, the authors consider the volume of messages posted about the same topic (i.e. using similar hashtags) and in a near geographical position. Therefore, small-scale events reported by a few number of users will not be detected, limiting the system to only identifying high-impact events.

The approach presented by Genc, Sakamoto and Nickerson (2011) employs Wikipedia articles for tweets categorization. Based on bag-of-words, they check for each word in the tweet (except stop words) whether there is a dedicated Wikipedia page for that word. The

topic of the tweets is given according to the category of the Wikipedia article selected. This approach can address distinct event types. However, even with the reduced length of the Twitter messages, analyzing every single word in the tweet can be very time-consuming. Therefore, term weighting techniques could be used to help to select the most relevant terms in the tweet, to be then searched in the Wikipedia platform.

3.2.2 Semantic Enrichment

In this category, we consider the works that employ LOD cloud datasets (e.g. DBpedia, Geonames, YAGO) and others knowledge bases (e.g. socialbakers). Each one of these knowledge sources describes resources (i.e. an entity in the dataset) using distinct properties belonging to specific vocabularies that can be retrieved (e.g. rdfs, skos), resulting in several concepts (i.e. semantic features) able to generalize and complement tweet contents.

The works developed by Schulz, Ristoski and Paulheim (2013) and Schulz, Guckelsberger and Janssen (2015) focus on small-scale incidents (e.g. car crashes, fires), which do not rely on massive amounts of tweets and bursts of vocabulary. In these works, enrichment is performed using properties retrieved from DBpedia, namely types (*rdfs:subClassOf*) and categories (*skos:broader*). Spatial, temporal and TF-IDF weighted terms were selected and evaluated as features, by constructing three different classifiers trained using car crash dataset (SCHULZ; RISTOSKI; PAULHEIM, 2013). In a more recent work (SCHULZ; GUCKELSBERGER; JANSSEN, 2015), the selected features were named entities related to textual features and advanced heuristics for identification of the presence of temporal and spatial expressions. The properties explored on DBpedia were *rdf:type* and *dct:subject*. An extensive analysis of the tweets datasets of similar incident types occurring in different cities and five different classifiers revealed that the features selected using NER tools were the most significant ones in building generalizable classification models for the target events. Although these authors point out that a huge amount of semantic features results from the enrichment can degrade the performance of event classifiers, this issue is not addressed in these works.

Rowe and Stankovic (2011) present an approach to align tweets to the events they report, by semantically enriching tweet contents to provide metadata that can be used to contextualize the event being analyzed. For the semantic enrichment step, the Zemanta web service was used to map the tweet content into concepts of the DBpedia knowledge base using the *dct:subject* property. To map tweets to a given event, two approaches were adopted: a) a proximity-based clustering approach, and b) supervised classification using the Naïve Bayes algorithm. The

authors also employed the SVM learning algorithm, but poor results were achieved due to the lack of discriminatory features.

3.2.3 Hybrid Enrichment

By hybrid enrichment, we refer to the approaches that combine at least two enrichment strategies: *external documents enrichment*, *semantic enrichment*, and *NER tools*. Due to the poor and sparse nature of tweets, important and complementary information might not be present in tweet contents, thus the *external documents enrichment* is used as a way to add additional information about the event analyzed. Regarding the *semantic enrichment*, its role is to expand the content of tweets through the generalization of specific terms and entities, resulting in a set of semantic features that better describe the domain of the event analyzed. The *NER tools* can be used to: a) extract named entities from the web documents to be used as complementary information; b) extract named entities from the tweet text and use them as input to the semantic enrichment, and c) extract named entities from the tweet text and use them as additional features to the classification step. Hence, by combining these strategies, we are incorporating related information, as well as generalizing the domain of the event analyzed, by adding useful knowledge about it.

Given a structured set of data about a specific planned event, the approach proposed by Packer et al. (2012) relies on using prior information, provided by an external document, and query expansion techniques as the base to extract concepts that are related to the type of the event analyzed. The paper is illustrated using prior information about a rock concert, which is assumed to be available as RDF triples. This external document is used as the source to expand the knowledge about the event. The set of terms extracted from the RDF structured data is used as input to perform semantic enrichment using the YAGO2 knowledge base. The label (i.e. *rdfs:label*) of the resulting resources are used as input to crawl the tweets related to the event. The temporal information of the messages recovered is used to determine whether the tweet is related to the object under analysis.

Abel et al. (2012b) propose a framework for filtering, searching and analyzing information about real-world incidents and crisis situations. The framework is connected to broadcasting emergency services, such that when a new incident occurs, the framework starts tracking information about the incident on Social Media. The broadcasting emergency service acts as an external document, in which NER tools are employed to retrieve categorical information about the incident (i.e. person involved and location). These named entities are then used as input to

the semantic enrichment using DBpedia. Specific properties of the resulting resources are used as input features to support the filtering of tweets that are related to the incident, as well as to provide data for summarizing and profiling incidents. The Twitcident framework employs two different approaches to decide whether a tweet is relevant or not for a specific incident. The first is a keyword-based approach, which is based on the information of the profile of the incident. The second one is a semantic filtering approach, which analyzes the similarity of the tweets and the profile using a vector space model representation and the Jaccard similarity measure. In the profile, tweets are categorized into reports about causalities, damages or risks, and also can describe the experience of the user on the event. The classification is a rule-based manner. The evaluation of the framework shows that the semantic enrichment boosts the performance of the filtering of tweets related to a specific incident.

The work developed by Li et al. (2016) aims to categorize tweets in real time. For the experiments, different tweet topic classification approaches were explored: a) weighted text model (i.e. employs tweet contents to construct the classifier); b) entity knowledge base (i.e. the knowledge base was built automatically by crawling data provided by Social Baker); and c) models based on word embedding (i.e. a distributed representation of words in a space with N dimensions), trained using Google News documents, general data collections (e.g. Reuters news articles, Wikipedia dump, UMBC web base corpus), and Twitter messages. The entity knowledge base was built considering specific entity categories (e.g. person, brand name, company, location) and the models based on word embeddings consider the semantic similarity among the words. These approaches were able to classify huge amount of tweets at real-time with high accuracy and to overcome the problems caused by sparseness in text classification:

- the models based on word embeddings provided the best performance, supporting the trend in the use of this type of model for text classification;

- to achieve real-time performance, Social Baker has been applied. This knowledge base meets this requirement because it is lightweight compared to the datasets available in the LOD cloud. Nevertheless, the content offered by this base is not as complete as those of the LOD cloud;

- as the baseline for the analysis, they rely on Open Calais categories, Latent Dirichlet Allocation (LDA) that uses a topic to represent a document or a sentence, and weighted text model.

3.3 Textual and Semantic Features for Event Identification and Classification in Tweets

Regardless the type of the enrichment employed, the features used as input and the kind of information extracted from the source are fundamental to the event classification performance.

3.3.1 Input Features

For specific features used as input to the contextual enrichment, the main approaches employed relies on:

- text extraction techniques (i.e. based on frequency or on the representativeness of the terms), to identify the textual features in the tweets. This kind of technique can also be used to recover URL mentioned in the post (VOSECKY et al., 2014; SANKARANARAYANAN et al., 2009);
- NER techniques to extract the entities that compose the tweet or the web document analyzed. The common named entities considered are *person*, *organization*, and *location* (SCHULZ; RISTOSKI; PAULHEIM, 2013; ABEL et al., 2012a; VOSECKY et al., 2014);
- extracting the elements *who*, *what*, *where*, and *when*, which are normally represented by named entities, nouns and verbs (FISICHELLA et al., 2011; LIU et al., 2016) and can also be represent by timestamps (VOSECKY et al., 2014).

The extraction and mapping of these features can be executed in different ways:

- two steps approach: the textual features (i.e. frequent or weighted terms, named entities) are extracted from the tweet text. These features are then used as input to identify the correspondent resource in the knowledge base (e.g. the term USA can be mapped to *dbpedia:United_States* in the DBpedia dataset) or to identify a specific Wikipedia article about that term/entity (e.g. the term USA can be linked to *https://en.wikipedia.org/wiki/United_States*) (ABEL et al., 2012b; PACKER et al., 2012; GENC; SAKAMOTO; NICKERSON, 2011). This is the commonly employed approach since the content to be enriched is delimited for features that are more likely to be related to the event domain;
- one step approach: the whole content of the tweet text is used to map specific resources in the knowledge base. In the works of Schulz et al. (2013, 2015), the whole tweet

text is submitted to the DBpedia Spotlight service[1] to identify and map named entities and terms to DBpedia resources. The resources recovered vary according to the context represented by the tweet. Then, the URI that represents each resource in the knowledge base is used to semantically enrich the Twitter message. However, this approach can produce a huge amount of semantic features, belonging to different domains, and which present no discriminative contribution to the event analyzed.

3.3.2 Output Features

Related work varies on the output features resulting from the enrichment task:

- Rosa et al. (2011) proposed to use all the textual content available in the web document to be incorporated into the tweet text;
- Vosecky et al. (2013) extracted named entities and TF-IDF weighted terms from web documents mentioned in the tweets;
- Genc, Sakamoto and Nickerson (2011) used the category information about a Wikipedia article to define the topic of the tweets analyzed;
- the object of the properties *skos:broader*, *rdfs:subclassOf*, *rdf:type*, and *dct:subject* available in the LOD cloud are the commonly concepts extracted and incorporated to the tweets (SCHULZ; RISTOSKI; PAULHEIM, 2013; ROWE; STANKOVIC, 2011; SCHULZ; GUCKELSBERGER; JANSSEN, 2015). Other works (ABEL et al., 2012a; PACKER et al., 2012) extract only the label of the resource (i.e. *rdfs:label*).

In addition to the features resulting from the contextual enrichment, other textual features can also be extracted from the tweet text to be used in the classification task, such as named entities (VOSECKY et al., 2014), spatial and temporal information (SCHULZ; GUCKELSBERGER; JANSSEN, 2015; VOSECKY et al., 2014; SAKAKI; OKAZAKI; MATSUO, 2010), and relevant terms using weighing techniques (e.g. TF-IDF) (VOSECKY et al., 2014).

[1] https://dbpedia-spotlight.github.io/demo/

3.4 Feature Incorporation

Related work proposes distinct approaches to incorporate the semantic features resulting from the contextual enrichment to the textual features of the tweet dataset (SAIF; HE; ALANI, 2012):

Replacement: replace all terms and entities recognized in the tweets by their correspondent semantic features.

Augmentation: maintain all terms and entities recognized in the tweets in addition to their semantic features.

Interpolation: it is a method able to interpolate an arbitrary type of feature (e.g. nouns, adjectives, semantics), considering the maximum likelihood estimation among the uni-grams from the original dataset and the interpolation component.

3.5 Algorithms for Event Classification

As mentioned, the events are classified using supervised learning techniques and clustering. The most common algorithms used to perform the classification task are Naïve Bayes and SVM (SAKAKI; OKAZAKI; MATSUO, 2010; ARAMAKI; MASKAWA; MORITA, 2011; SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; ROWE; STANKOVIC, 2011; SANKARANARAYANAN et al., 2009). The use of Random Forest, J48, and Ripper rule learner (JRip) is also reported (SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015). Clustering is employed in the situations where topic classification is required, where the most common techniques are Latent Semantic Analysis (LSA), LDA, and K-means (VOSECKY et al., 2013; GENC; SAKAMOTO; NICKERSON, 2011; ROSA et al., 2011).

3.6 Final Remarks

Tables 3.1 and 3.2 summarize the main aspects discussed in this Chapter regarding related work.

Considering the works presented in Tables 3.1 and 3.2, the classification approach can be based only on the tweet content and words arrangement (PETROVIC et al., 2013; SAKAKI;

OKAZAKI; MATSUO, 2010; ARAMAKI; MASKAWA; MORITA, 2011; LIU et al., 2016), or other enrichment techniques can be applied: a) by extracting information from web documents related to the event analyzed (VOSECKY et al., 2014; SANKARANARAYANAN et al., 2009; ROSA et al., 2011; GENC; SAKAMOTO; NICKERSON, 2011); b) by exploring the properties available in the knowledge bases, such as DBpedia and YAGO2 (SCHULZ; RISTOSKI; PAULHEIM, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; ROWE; STANKOVIC, 2011); c) by combing different strategies to generalize the tweet content (PACKER et al., 2012; ABEL et al., 2012b; LI et al., 2016).

Different features are selected to be used in the event classification process, such as named entities (SCHULZ; RISTOSKI; PAULHEIM, 2013; VOSECKY et al., 2014; ABEL et al., 2012b; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; FISICHELLA et al., 2011; PACKER et al., 2012; LIU et al., 2016), spatial and temporal information (SCHULZ; GUCKELSBERGER; JANSSEN, 2015; VOSECKY et al., 2014), relevant terms using weighing techniques (e.g. TF-IDF) (SAKAKI; OKAZAKI; MATSUO, 2010; SCHULZ; RISTOSKI; PAULHEIM, 2013; SANKARANARAYANAN et al., 2009) or manually selected terms (PACKER et al., 2012). Related work also differs in the semantics and properties extracted from the resources explored (i.e. web documents, sources related to NER tools or LOD cloud). Moreover, only a few works reported about using feature selection techniques to improve the classification performance (ROWE; STANKOVIC, 2011).

Leveraging these related work experiences, the framework proposed in this work was designed as follows. The textual features defined are: a) *Vocabulary*; b) *Agents*; and c) *Location*. To provide context to tweets, we propose a process to enrich tweet texts with contextual information that combines these two complementary approaches: a) *external documents enrichment* using related web pages for extending the conceptual features contained within the tweets; and b) *semantic enrichment* using the LOD cloud to add related semantic features. Regarding the knowledge base, most of the works performed the semantic enrichment using the DBpedia dataset, which supports our choice of employing it in our approach. For text classification, algorithms such as SVM and Naïve Bayes are widely used, which motivates us to employ the same algorithms in the classification step.

Table 3.1: Summary of related work

Work	Description	Features			Selection of features	Learning Technique	Event Type
		Tweet	External	Semantic			
(PETROVIC et al., 2013)	Examine whether the Twitter can overlap the traditional news media in the faster disclosure of breaking news	Unspecified	NO	NO	NO	KNN and similarity	General (unspecified, planned, and unplanned)
(ARAMAKI; MASKAWA; MORITA, 2011)	An approach for detecting influenza epidemics	BOW	NO	NO	NO	SVM	Epidemics (specified, unplanned)
(SAKAKI; OKAZAKI; MATSUO, 2010)	Detection and Classification of natural disasters (earthquake)	Domain representative keywords and context words	NO	NO	NO	SVM	Natural disasters (specified, unplanned)
(LIU et al., 2016)	System to discover breaking news faster than the traditional mews media sources	Named entities, nouns and verbs, which represent the elements: who, where, and what.	NO	NO	NO	Clustering (own implementation)	General (unspecified, planned, and unplanned)
(VOSECKY et al., 2014)	An approach for topic modeling	Named entities (person, organization, and location) and timestamps	Analysis of the URL content	NO	NO	K-means, DBSCAN, Single-pass Incremental Clusterer, and Direct	General (unspecified, planned, and unplanned)
(SANKARANARAYANAN et al., 2009)	System for breaking news identification and cluster on Twitter	Hashtags, geographical position, and TF-IDF weighted keywords	Analysis of the URL content	NO	NO	Naive Bayes and clustering	General (unspecified, planned, and unplanned)
(ROSA et al., 2011)	A URL-based approach to automatically clustering and classifying tweets	Hashtags and TF-IDF	Analysis of the URL content	NO	NO	LDA, K-Means, and Rocchio	General (unspecified, planned, and unplanned)

Source: the author.

Table 3.2: Summary of related work (cont.)

Work	Description	Features			Selection of features	Learning Technique	Event Type
		Tweet	External	Semantic			
(GENC; SAKAMOTO; NICKERSON, 2011)	Map the tweet content to its most similar Wikipedia article, for topic categorization	BOW	Wikipedia	NO	NO	LSA	General (unspecified, planned, and unplanned)
(ROWE; STANKOVIC, 2011)	An approach to align the tweets to events they report	Named entities (Zemanta)	NO	DBpedia (dct:subject, rdf:type)	YES	Proximity-based clustering approach and Naive Bayes	General (unspecified, planned, and unplanned)
(SCHULZ; RISTOSKI; PAULHEIM, 2013)	Detection and Classification of small-scale incidents	TF-IDF, spatial and temporal expression, named entities	NO	DBpedia (rdf:subClassOf, skos:broader)	NO	SVM, Naive Bayes, and JRip	Incidents (specified, unplanned)
(SCHULZ; GUCKELSBERGER; JANSSEN, 2015)	Detection and Classification of small-scale incidents	Named entities, temporal and spatial expression	NO	DBpedia (rdf:type, dct:subject)	NO	J48, Naive Bayes, JRip, LibLinear (SVM), Random Forest	Incidents (specified, unplanned)
(PACKER et al., 2012)	Detecting tweets related to a rock concert	Named entities	Concert program	YAGO2	NO	Pearson's Correlation Coefficient	Rock concert (specified, planned)
(ABEL et al., 2012b)	Filtering of the tweets employing a keyword-based approach and a semantic filtering approach, using a vector space model representation	Named entities (person, location, and organization)	Broadcasting emergency service and the analysis of the URL content	DBpedia (rdfs:label)	NO	Handcrafted rules (Jaccard similarity to detect the relevant tweets)	Emergency situations (specified, unplanned)
(LI et al., 2016)	Present several approaches to categorize tweets in real time	Unspecified	Google News and other general data collections	Socialbaker	NO	SVM	General (unspecified, planned, and unplanned)

Source: the author.

4 A FRAMEWORK FOR EVENT CLASSIFICATION IN TWEETS BASED ON HYBRID SEMANTIC ENRICHMENT

This chapter presents the hybrid semantic enrichment framework proposed in this work. First, an overview of the architecture of the framework is presented, and then each step of the process is detailed.

4.1 Overview

As mentioned in the previous chapter, proposed approaches for event classification vary in the features selected, the enrichment technique employed, as well as in the classification algorithm applied. Often these choices take into account specific assumptions about the event type and the goal of the proposed application. Thus, it is difficult to reproduce these approaches to another event type and directly compare them to other applications.

We aim at leveraging related work and previous experiences on contextual enrichment to build a generalizable framework for event classification in tweets, which is able to deal with specified events of distinct natures. Thus, one will be able to employ the framework and compare its performance for different event types, being useful also to compose baselines for other applications and classification techniques.

The proposed event classification framework considers the following main elements:

- conceptual features: by analyzing several event definitions, we identified a set of features that are commonly used to characterize distinct types of events. Based on these definitions and to achieve our goal of classifying events of different natures, we defined a set of core features to be used in the framework: a) *Vocabulary*, i.e. terms that represent the topic or subject of the event; b) *Agents* that are involved or affected by the event; and c) *Location* that represents geographic information about the event;

- contextual enrichment: given the distinct enrichment techniques employed in related work, we propose a hybrid enrichment process by combining: a) external documents enrichment to use the content of related web documents as a way to obtain more information about the event, and to overcome the poor and sparse textual content of tweets; b) NER tools to recover specific conceptual features from the tweet text and web documents; and c) resources available in the LOD cloud to semantically enrich the set of

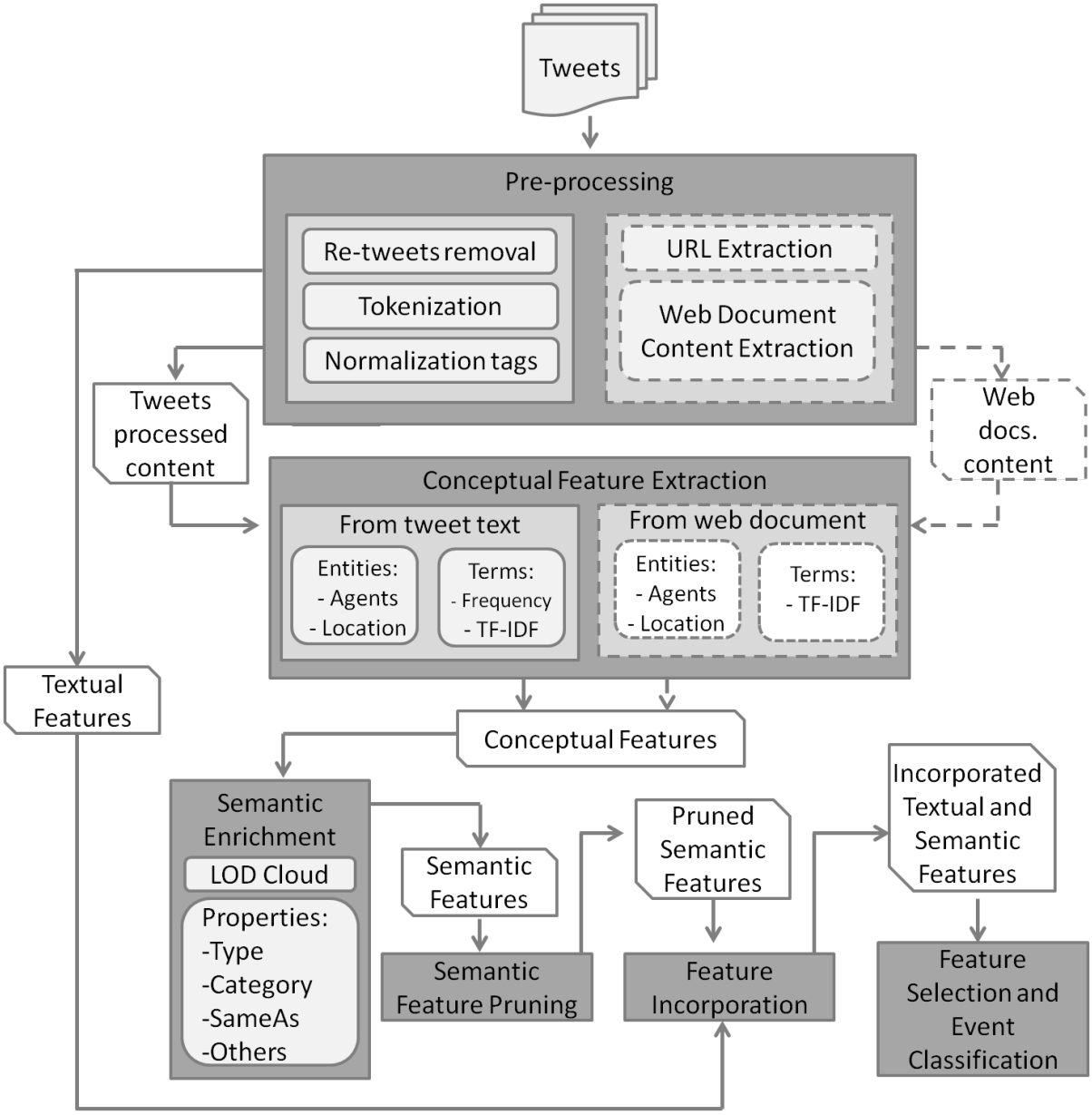
conceptual features, allowing the generalization of the event domain and improving the event classification process;

- selection of relevant semantic features: good results in the classification task are dependent on discriminatory features in the training dataset. Given the Hybrid Semantic Enrichment process executed, general-purpose feature selection techniques may be enough to handle the sparse set of semantic features resulting from this process, and select the most relevant ones, for the target event. As an attempt to solve this problem, we propose a specific-purpose pruning algorithm, to select the most discriminative semantic features resulting from semantic enrichment. This pruning algorithm is complementary to general-purpose feature selection techniques.

Figure 4.1 presents the proposed framework for event classification in tweets. It 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 semantic 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, the resulting dataset is submitted as input to the *classification* step, possibly preceded by the deployment of a *feature selection* method. Details on these steps are provided in the remaining of this chapter. Figure 4.2 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.

[1] http://cmci.colorado.edu/ mpaul/downloads/flu_data.php

Figure 4.1: Framework for event classification in tweets based on Hybrid Semantic Enrichment



Source: the author.

4.2 Pre-processing

Given a set of tweets as input, the goal of this step is to perform basic pre-processing actions, such as tokenization, re-tweets removal to avoid redundancy (SCHULZ; GUCKELSBERGER; JANSSEN, 2015), and the normalization of specific features (i.e. @User, URLs, and emoticons to T_USER, T_URL, and T_EMOT) (MEDVET; BARTOLI, 2012). These data preparation techniques reduce the number of features in the dataset and avoid over-fitting problems by allowing more generalized models.

The recognition of URLs mentioned in the tweets and the extraction of the textual con-

Figure 4.2: A running example according to the proposed framework

<table>
<tr><td>Original tweet</td><td colspan="3">@user: Officials worry about swine flu preparedness amid budget cuts. http://bit.ly/J2gZX</td></tr>
<tr><td>Web document content</td><td colspan="3">http://bit.ly/J2gZX: “... six years of worrying about bird flu did much to prepare the United States for the current swine flu outbreak.”</td></tr>
<tr><td>Processed tweet</td><td colspan="3">T_USER: Officials worry about swine flu preparedness amid budget cuts T_URL</td></tr>
<tr><td rowspan="2">Conceptual Features Extracted</td><td colspan="2">From tweet text</td><td>From URL</td></tr>
<tr><td colspan="2">Swine (frequent/representative terms)
Worry (representative term)</td><td>Flu (representative term)
United States (location)</td></tr>
<tr><td colspan="2">Semantic Features (rdf: type)</td><td colspan="2">Semantic Feature Pruning</td></tr>
<tr><td colspan="2" rowspan="2">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</td><td>Too Generic</td><td>Too Specific</td></tr>
<tr><td>owl:Thing</td><td>wikidata:Q12136
umbel-rc:AilmentCondition
yago:WikicatEnglish-speakingCountriesAndTerritories</td></tr>
<tr><td>Pruned Semantic Features</td><td colspan="3">dbo:Disease, dbo:PopulatedPlace</td></tr>
<tr><td colspan="4">Incorporated textual and semantic features</td></tr>
<tr><td colspan="4">T_USER, officials, worry, about, swine, flu, preparedness, amid, budget, cuts, T_URL, dbo:Disease, dbo:PopulatedPlace, ...</td></tr>
</table>

Source: the author.

tents of related web documents are also executed in this step, as presented in Figure 4.1. This content will be used as input to the Conceptual Feature Extraction step. The outputs of this step are:

- the tweet uni-grams (i.e. textual features), including the ones that represent user annotation (i.e. @User), URLs, and emoticons converted to the symbols T_USER, T_URL, and T_EMOT, respectively;
- the pre-processed tweet content for conceptual feature extraction;
- a list of the URLs extracted from the tweet text and its respective content.

The running example on Figure 4.2, shows the pre-processed tweet content and the web document content extracted.

4.3 Conceptual Feature Extraction

The goal of this step is to extract a set of conceptual features from the (pre-processed) documents resulting from the previous step. Conceptual features are the ones to be semantically enriched in the next step of the process.

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. Those issues are discussed in the remaining of this section.

4.3.1 Core Feature Types

Based on *Definition 2.1* proposed in Chapter 2 (i.e. *"an event is an occurrence, represented by a topic, that occurs in a specific time and can involve one or more locations and agents"*) and by analyzing the importance of each feature described by the event definitions in Table 2.1, we defined a set of core features to represent the common ones useful for characterizing events in general (ROMERO; BECKER, 2016a):

- *Vocabulary* refers to terms that are frequent, representative of a domain, or created in bursts for a specific event (BECKER; NAAMAN; GRAVANO, 2011; MCMINN; MOSHFEGHI; JOSE, 2013; MEDVET; BARTOLI, 2012; SAKAKI; OKAZAKI; MATSUO, 2010; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; ANANTHARAM et al., 2015). This set of terms describes 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. Related work also refers to the subject or topic of an event as the element *what* of an event reported (LIU et al., 2016; FISICHELLA et al., 2011);

- *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, a Politician in a debate, 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; OKAZAKI; MATSUO, 2010). Agents are also referenced as the element *who* (LIU et al., 2016; FISICHELLA et al., 2011);

- *Location* is an important type of feature, which can either refer to a geographic property that characterizes the event itself or to a place in which the event happened or from which

it is reported (ANANTHARAM et al., 2015; BECKER; NAAMAN; GRAVANO, 2011; MCMINN; MOSHFEGHI; JOSE, 2013; SCHULZ; GUCKELSBERGER; JANSSEN, 2015; SAKAKI; OKAZAKI; MATSUO, 2010). Considering event detection approaches based on users as sensors, the location of the report or the user is quite relevant, and it is synonym of the location of the event (SAKAKI; OKAZAKI; MATSUO, 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; SCHMIDT; STRUFE, 2015). Therefore, location information needs to be identified in the text itself or from the user profile. In the related work (LIU et al., 2016; FISICHELLA et al., 2011), location is also referenced as the element *where*.

Scale and the temporal component are also very important, but were not considered in this work because we are dealing with the task of improving event classification. By delimiting the scale, we can restrict the type of event to be analyzed by the framework. Regarding the temporal component, it will be explored in future work to produce a correlation between similar events that occurred in a different period of time (e.g. the London Olympics in 2012 and the Rio Olympics in 2016).

We developed initial experiments reported in Appendix A to analyze whether a type of core feature played a more specific role according to the nature of the event (e.g. sportive, natural disaster). We could not observe a pattern, as the results showed that each core feature can help event classification in different ways and no single combination provides the best results for all kinds of events (ROMERO; BECKER, 2016a). Thus, we opted to extract all the core features from the dataset, independent of the event type analyzed.

4.3.2 Source

As aforementioned, tweet contents present a poor and informal vocabulary, which cannot be enough to characterize the event as a whole, limiting the identification of relevant information about the event analyzed.

As an attempt to overcome this problem, in addition to the tweet text, the conceptual features can also be extracted from external documents, as a way to complement the description of the event analyzed and better represent its domain. To select these external documents different techniques can be applied, such as the recognition of Wikipedia articles that describe specific terms related to the event, the recognition of news sites and blogs that reported the event, the

extraction of URLs mentioned in the tweet, among other (SANKARANARAYANAN et al., 2009; VOSECKY et al., 2014; GENC; SAKAMOTO; NICKERSON, 2011).

In this work, we opted to extract the conceptual features from both (i.e. the tweet text and the web documents mentioned in the tweets). These related web documents are recognized through the URLs mentioned in the post and can represent another tweet or some web page (e.g. sportive news, blog).

4.3.3 Extraction Techniques

Regarding the tools and techniques used to extract the core features, we propose to identify Agents and Location using NER tools, since this kind of tool is the most indicated to recognize specific categories of terms presented in the text (SAIF; HE; ALANI, 2012; ABEL et al., 2012a; SCHULZ; RISTOSKI; PAULHEIM, 2013).

By recognizing the named entities presented in the text, we are able to recover a greater number of tweets related to the domain of the event analyzed, thus improving Recall, which is the focus of the contextual enrichment approach.

Several NER tools are available, some of them using private knowledge bases, while others are based on open source bases. We performed experiments comparing two widely used NER tools (SAIF; HE; ALANI, 2012), namely Open Calais[2] and AlchemyAPI[3], in preliminary experiments. Open Calais was able to recognize a greater number of named entities in the datasets used as input compared to AlchemyAPI. Thus, this tool was employed in our experiments.

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. The latter is extracted using term weighting techniques, such as TF-IDF given a threshold (SCHULZ; RISTOSKI; PAULHEIM, 2013).

From the tweet content we extract:

- the named entities representing *Agents* and *Locations*;
- the top-k frequent terms;
- the representative terms using TF-IDF weighting.

[2]http://www.opencalais.com/
[3]http://www.alchemyapi.com/

From the web documents, we decided to extract:

- the named entities representing *Agents* and *Locations*;
- the representative terms using the TF-IDF weighting technique.

After some preliminary tests, we decided not to extract the top-k frequent terms from external web documents, due to the characteristics of this type of document, generally blogs and news sites, in which the resulting list presented terms that were not necessarily related to the domain of the event analyzed. In addition, these experiments also demonstrated that representative terms extracted using TF-IDF subsumed the domain-related frequent terms.

The example of Figure 4.2 shows the entities/vocabulary extracted from the tweet text and the referenced web page. Note that the vocabulary extraction strategy does not select a feature from the tweet text if it contributes to sparsity (e.g. "budget", which is neither frequent or representative). On the other hand, 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 both tweets texts and external web documents helps us to identify features that better represent the domain of the event analyzed.

4.4 Semantic Enrichment

Given a set of conceptual features as input, the aim of this step is generalizing these features to obtain more representative domain concepts, by retrieving associated knowledge from the LOD cloud. The resulting output is a set of semantic features.

The Semantic Enrichment step involves two tasks:

- mapping the conceptual feature into a resource in a specific knowledge base in the LOD cloud;
- retrieving properties of the resource described in the knowledge base. Different properties can be selected to assign meaning to tweets, such as *Type*, *Category* and *sameAs*.

Likewise, according to the application purpose, different knowledge bases can be deployed, such as DBPedia, Geonames, YAGO, etc (PACKER et al., 2012; SCHULZ; GUCKELSBERGER; JANSSEN, 2015). In this work, 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. The object of the *rdf:type* property (i.e. also called in this work as concepts or semantic features) is used as a complementary information, which will help us to generalize the knowledge about the conceptual feature. For example, by analyzing the object of the *rdf:type* property in the DBpedia knowledge base for the conceptual features Didier Drogba and Ramires, we can observe that they are related to similar concepts (e.g. *dbo:SoccerPlayer*, *yago:Athlete109820263*), which may help in the Classification step for soccer or sportive events.

The matching between the conceptual features and the resource in the knowledge base can be performed automatically (e.g. the *DBPedia Spotlight* operator from the RapidMiner platform), or by an *ad hoc* method (e.g. SPARQL queries). Figure 4.2 illustrates the semantic enrichment for our running example using *rdf:type* property and DBpedia.

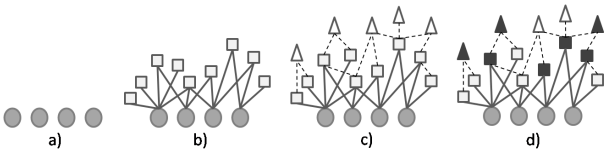
4.5 Semantic Feature Pruning

This step aims to reduce the volume of semantic features resulting from the previous one. For that purpose, 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 or too specific. This pruning phase is applied over the semantic features only, and the result is a reduced set of semantic features to be incorporated into the textual features from the dataset.

Considering the running example of Figure 4.2, 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*). The addition of these features to the training dataset 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. This step aims at selecting the relevant semantic features that add the proper level of generalization to the conceptual features.

Next, we describe the pruning algorithm operation and the technique proposed for automatically define the pruning threshold value.

Figure 4.3: Pruning concepts



Source: the author.

4.5.1 Algorithm Description

The PageRank method was originally developed for rating web documents based on the link structure of the Web graph (PAGE et al., 1999). Using forward links (*outedges*) and backlinks (*inedges*), a random surfer visits these pages computing their salience in the graph. Pages with high scores are regarded as representative pages, 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. Figure 4.3 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 semantic properties (i.e. *rdf:type*);

c) these nodes are interconnect using super/subclass relationships also retrieved from the LOD cloud using additional SPARQL queries;

d) given the built graph, PageRank is used to calculate the salience scores. Then, specific thresholds are applied to remove the nodes above/under pruning thresholds, 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, represented as an adjacency matrix, using the resources, their respective concepts, sub/superclass relationships between these concepts (i.e. *rdfs:SubClassOf*), as well as the number of sub/superclasses. In

Algorithm 1 PageRank-based feature pruning

1: **Input:** *Concepts* = concepts extracted from the LOD cloud
2: **Output:** Ø= set of relevant concepts
3: *Property* ← **getTypes**(Concepts)
4: *SubClass* ← **getSubClassOf**(Property)
5: *CountSC* ← **getCountSubClassOf**(SubClass)
6: *Matrix* ← **getAdjacencyMatrix**(SubClass)
7: *PRgraph* ← **calculatePageRank**(Matrix)
8: Ø← **performPruning**(*CountSC*, *PRgraph*)
return Ø

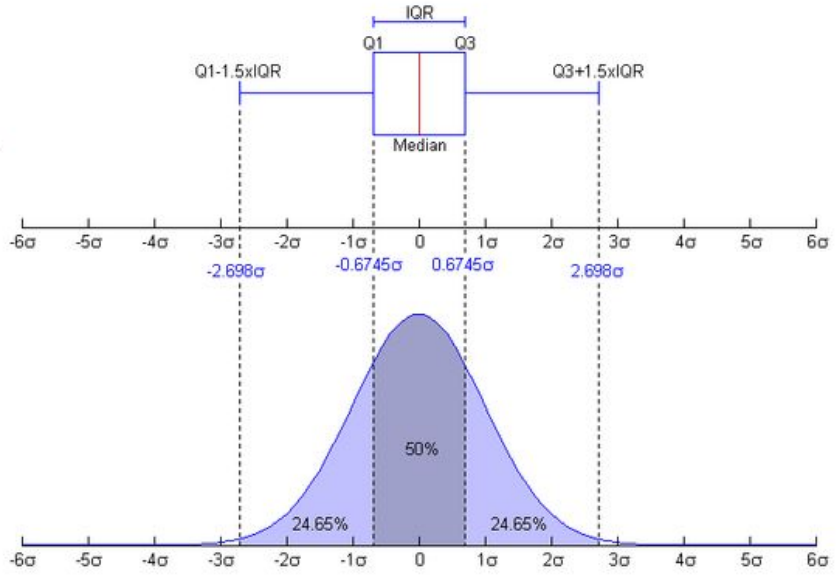
line 7, the scores are calculated following the traditional PageRank technique to build the graph: a) each node starts with a default score; b) we calculate the score of each node based on the random surfer visit and the amount of backlinks to the node; c) if the random surfer visits a node without forward links, we apply the damping factor and jumps to another page. We devised two strategies to automatically defining the pruning upper/lower thresholds based on the distribution of salience scores and remove the semantic features that are too generic/specific (line 8), which are detailed in the Section 4.5.2.

As output, this step results in a set of semantic features that better represent the event analyzed. The set of relevant semantic features is composed of the semantic features retrieved in LOD cloud dataset by analyzing a specific property, in addition to the other classes identified by the sub/superclass relationship. In the example of Figure 4.2, we would exclude *owl:Thing*, as a generic feature, as well as specific features, such as *umbel-rc:AilmentCondition*.

4.5.2 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 a deep 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), as presented in Figure 4.4. The rationale behind it is that by using the median and the equations based on it, we are able to deal with these distorted values, since the median is a measure that is not influenced by very larger or very small values.

Figure 4.4: Quartiles representation



Source: http://slideplayer.it/slide/8246/.

Three strategies to automatically defining the pruning thresholds were considered:

- the values of the upper and lower quartiles of the PageRank salience scores distribution. This strategy produced very poor results, and was discarded;
- the values of the upper and lower quartiles of the PageRank salience scores distribution, where the former is adjusted using the IQR (Equation 4.1). This strategy is referred to as IQR. The pseudo-code is listed in Algorithm 2, where Qs and Qi represent the upper and lower quartiles, respectively;
- the values of the upper/lower quartiles of both the PageRank salience scores distribution, and the number of sub/super class distribution. This strategy is referred to as QUARTILES, and the pseudo-code is listed in Algorithm 3.

The rationale for the IQR strategy is to identify concepts that are too generic as an approximation of outliers. We also attempted to spot too specific concepts using the minimum value, but our experiments revealed this value was too inclusive, allowing the selection of over-specialized concepts.

$$IQR = Qs - Qi \tag{4.1}$$

Algorithm 2 Summary of the IQR strategy

```
1: Input: PRgraph = graph of the concepts and its scores
2: Output: Ø= set of relevant concepts
3: QuartilesPR ← calculateQuartile(PRgraph)
4: IQR ← QuartilesPR.Qs − QuartilesPR.Qi
5: MaxValue ← (QuartilesPR.Qs + (IQR ∗ 1.5))
6: LowerValue ← QuartilesPR.Qi
7: for i in PRgraph do
8:     PRvalue ← PRgraph[i].value
9:     if (PRvalue > LowerValue AND PRvalue < MaxValue ) then
10:        Ø← PRgraph[i] −
   return Ø
```

The QUARTILES strategy was devised because 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.

Algorithm 3 Summary of the QUARTILES strategy

```
1: Input: PRgraph = graph of the concepts and its scores, CountSC = number of sub/super
   class of each concept
2: Output: Ø= set of relevant concepts
3: QuartilesPR ← calculateQuartile(PRgraph)
4: QuartilesSC ← calculateQuartile(CountSC)
5: for i in PRgraph do
6:     PRvalue ← PRgraph[i].value
7:     SCvalue ← CountSC[i].value
8:     if (PRvalue > QuartilesPR.Qi AND PRvalue < QuartilesPR.Qs ) then
9:         if (SCvalue > QuartilesSC.Qi AND SCvalue < QuartilesSC.Qs ) then
10:            Ø← PRgraph[i] −
   return Ø
```

We performed a detailed analysis of the IQR and QUARTILES, where the latter produced slightly better results. Therefore, it was adopted in the experiments reported in Chapter 5. The performance of the IQR strategy is detailed in Appendix B, together with a comparison with QUARTILES.

We also observed that combining this strategy with general-purpose feature selection algorithms, such as *CfsSubsetEval* and *InformationGain*, can produce even better results, since

it considers the whole dataset, instead of only the semantic features. Details about these experiments are reported in Appendix B.

4.6 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; HE; ALANI, 2012) because we are able to maintain the original conceptual feature and additionally to include generalized information about it. Thus, the original conceptual features will contribute to the Classification step as a textual feature and the concepts from the LOD cloud as the semantic features, which together can improve the results of the event classification process.

The training dataset contains both the textual tokens and the pruned semantic features extracted from the LOD cloud. As presented in the running example in Figure 4.2, the textual tokens extracted from the tweet text are incorporated with the pruned semantic features resulting from the previous step.

4.7 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.

Given the several feature selection algorithms that can be applied, we selected two of them to perform our experiments, namely *CfsSubsetEval* and *InformationGain*. The former considers the correlation between the features and the classes, and the latter consider the information gain produced for each feature to define the most relevant ones.

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). As the output of this step, we have an event classification model.

4.8 Final Remarks

In this chapter, we discussed the proposed framework for event classification in tweets. The framework is composed of six steps, in which different tools and techniques are employed. We leverage different techniques proposed in related work and integrated them in a unifying framework. In the next chapter, we evaluate the contribution of the key steps in the classification of event related tweets.

5 EVALUATION EXPERIMENTS

In this chapter, we describe the two main experiments performed to evaluate the contribution of the proposed framework for event classification in tweets. The first one (*Experiment #1*) aims at comparing the performance of the event classification process without any kind of contextual enrichment against the use of the proposed hybrid semantic enrichment framework, as well as analyze the performance of the PageRank-based pruning algorithm proposed. The second one (*Experiment #2*) aims at evaluating the performance of the proposed framework against an approach based on word embeddings, which has been widely used in NLP applications, due to the good results produced in other applications. Besides that, a word embeddings based approach considers the semantic similarity among the words, which can be treated as a primitive form of contextual enrichment. Other experiments are described in the Appendices.

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 traditional feature selection algorithm;
- the use of conceptual features from external web documents, in addition to concepts extracted from tweet text only;
- the generality of the framework by considering events of distinct natures;
- the semantic enrichment compared to a word embeddings based approach.

In the remaining of this chapter, we describe the general setups employed for both experiments and the results achieved in each one of them.

5.1 Target Event Datasets

We performed the experiments using seven target event datasets of distinct natures. Table 5.1 presents the name of the datasets used as the target, the number of tweets used from each dataset, the event type that each dataset corresponds to (i.e. planned events, natural disaster, and commemorative dates), and the number of tweets in the resulting dataset after randomly select the tweets from other target events. Details about all datasets employed are presented below:

- *FA Cup Final (FaCup)*: this dataset includes 18,000 tweets crawled during the final of the football season in England in 2012, using a specific set of keywords selected by experts (AIELLO et al., 2013);

- *2012 Olympic Games (Olympics)*: tweets related to the women's gymnastics competition at the London Olympic Games (**??**), filtered using specific hashtags. This dataset is composed of about 1,000 tweets;

- *Halloween*: this dataset is composed of tweets extracted from the streaming Twitter API using the keyword "Halloween". The posts were collected in the period of 10/29/2015 to 11/01/2015. We retrieved about 900,000 tweets;

- *Hurricane Sandy (HSandy)*: this dataset[1] was built by crawling tweets during the occurrence of the Hurricane Sandy that hit New York in 2012, using specific hashtags. It consists of 4,085 tweets manually annotated as relevant and irrelevant. We used from this dataset just the tweets annotated as relevant to represent a natural disaster situation;

- *Alberta Floods*: tweets related to the floods that hit Alberta, Canada, in 2013 (OLTEANU; VIEWEG; CASTILLO, 2015). The dataset is composed of about 1,000 annotated tweets;

- *Australia bushfire*: tweets crawled in 2013 during the Australia wildfires (OLTEANU; VIEWEG; CASTILLO, 2015). This dataset has about 1,200 tweets;

- *Influenza*: a collection of 2 billion tweets crawled from May 2009 to October 2010, and 1.8 billion tweets collected from August 2011 to November 2012. The tweets were annotated as concerned awareness, infection, media or unrelated (LAMB; PAUL; DREDZE, 2013). This dataset has about 1,400 annotated tweets[2], from which we used the ones annotated as concerned awareness and infection;

- *SemEval-2016*: this dataset was made available for the SemEval-2016 Task 4[3] for Sentiment Analysis purposes, and contains about 10,000 tweets annotated as positive, negative or neutral. Tweets from this dataset were used to compose the negative example in our experiments.

The positive events were extracted from the corresponding event dataset, and the negative ones were randomly selected from the other datasets. We also used the data available

[1] https://github.com/pavan046/benchmark-events-tweets-dataset
[2] http://cmci.colorado.edu/ mpaul/downloads/flu_data.php
[3] http://alt.qcri.org/semeval2016/task4/

Table 5.1: Description of the target datasets

Dataset	# of tweets used	Event type	# resulting dataset
FaCup	1502	Sportive - Planned event	4502
Olympics	1036	Sportive - Planned event	4036
Halloween	1551	Commemorative date - Planned event	4551
HSandy	1516	Natural disaster - Unplanned event	4516
Alberta Flood	950	Natural disaster - Unplanned event	2868
Australia Bushfire	881	Natural disaster - Unplanned event	2692
Influenza	1380	Epidemic - Unplanned event	4257

Source: the author.

in the SemEval-2016 Task 4, to simulate the existence of tweets not related to any particular event or subject target. 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. To produce uniform datasets that do not undermine the analysis due to the discrepancy in the number of positive examples, we limited the positive event tweets to 1,600. For example, the Olympics dataset is composed of all posts from 2012 Olympic Games dataset as the positive label (i.e. 1036 tweets), and a set of tweets randomly selected from 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.

5.2 Baselines

Given the distinct goals of our experiments, we defined different baselines to be used in each one of them:

- *Experiment #1*: the baseline is composed by all alphabetic uni-grams extracted from each tweet dataset, including the normalized symbols (i.e. T_USER, T_EMOT, and T_URL). 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;

- *Experiment #2*: we adopted as the baseline a word embeddings based approach, which considers the co-occurrence of the words from the vocabulary encoded as real-valued vectors in a dimensional space. We choose this approach since it captures the syntactic and semantic characteristics of a word, allowing their representation in different contexts and is considered as a primitive form of contextual enrichment. Thus, it allows us to compare our hybrid enrichment approach to another alternative for contextual enrichment.

For the experiments, we adopted the pre-trained word vectors using GloVe[4], which was produced over 2 billions tweets, representing a 1.2 million vocabulary. We employed the GloVe model built in a 100-dimension space (i.e. each word is associated to a vector with 100 numbers that are semantically similar to them, given different contexts explored in the corpus). Regarding the word features, in our experiments, we applied the mean technique (LIU et al., 2015), in which we calculate the mean of all word vectors recognized for each term in a tweet. Then, we trained the model using the supervised algorithms NB and SVM in a 10-fold cross-validation configuration.

5.3 Experiment #1: No Contextual Enrichment vs. Hybrid Semantic Enrichment

In the first experiment, we compared the benefits of using the proposed hybrid semantic enrichment framework, against the use of the classification of event-related tweets based solely on textual features, i.e. no use of any contextual enrichment technique. We also analyzed the contribution of the different steps of the framework and how they collaborate to the improvement of the final results.

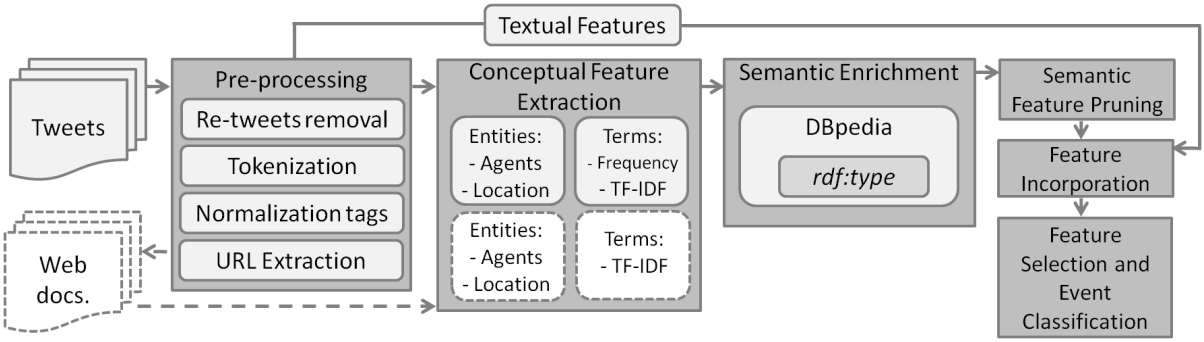
5.3.1 Preliminary Experiments

At an early stage of this research, we developed preliminary experiments to analyze the contribution of each type of core feature in the classification of distinct event types (ROMERO; BECKER, 2016a). These experiments involved four of the datasets described in Section 5.1 (i.e. FaCup, Olympics, Halloween, and HSandy datasets), the extraction of the contents of tweets only, semantic enrichment using DBpedia and the *rdf:type* property, general purpose feature selection algorithms and two classification algorithms. The main results of these experiments are reported in Appendix A.

The main conclusions were: a) each type of core feature help event classification in different ways, but no single combination provides the best results for all kinds of events; b) DBpedia showed a good coverage (i.e. 92% of the concepts extracted), but the semantic enrichment yielded quite sparse datasets, with too many features; c) the improvements provided by the adoption of general-purpose feature selection algorithms were dependent on the classification algorithm. These preliminary results motivated the improvement of the semantic enrichment

[4]http://nlp.stanford.edu/projects/glove/

Figure 5.1: Summarized pipeline of the Event Classification process



Source: the author.

framework by including external source enrichment and a domain-specific semantic feature pruning technique, as described in Chapter 4. Due to the lack of patterns, we decided to combine the core features into two groups for evaluation purposes, namely: a) the named entities composed of the Agents and Locations, referred from now on to as *NER*, and b) the frequent and representative vocabulary, referred to as *TERMS*.

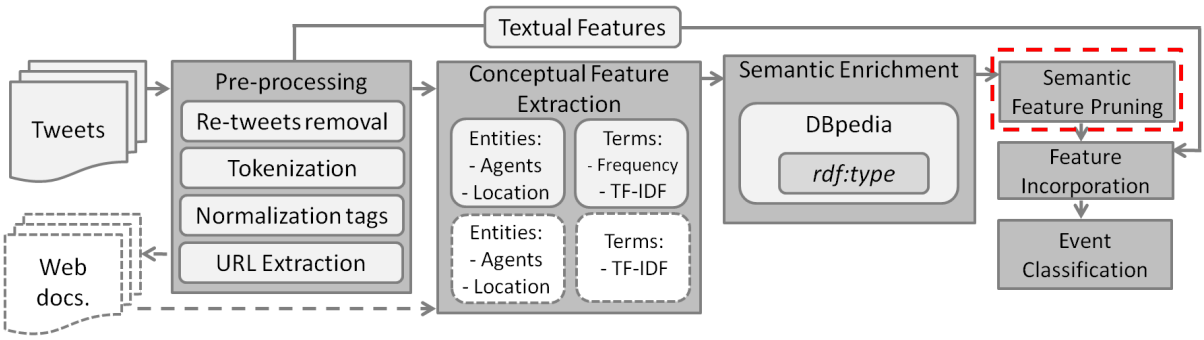
5.3.2 Experiment Description

Based on the core features defined in Chapter 4 and according to our preliminary experiments (i.e. Appendix A), we prepared the following mining datasets for each target event to further analyze the contribution of the types of features proposed:

- NER: contains the uni-grams extracted from the tweets, incorporated with the semantic features resulting from the enrichment of agent and location entities;
- TERMS: contains the uni-grams extracted from the tweets, incorporated with the semantic features resulting from the enrichment of frequent and representative terms;
- ALL: 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.

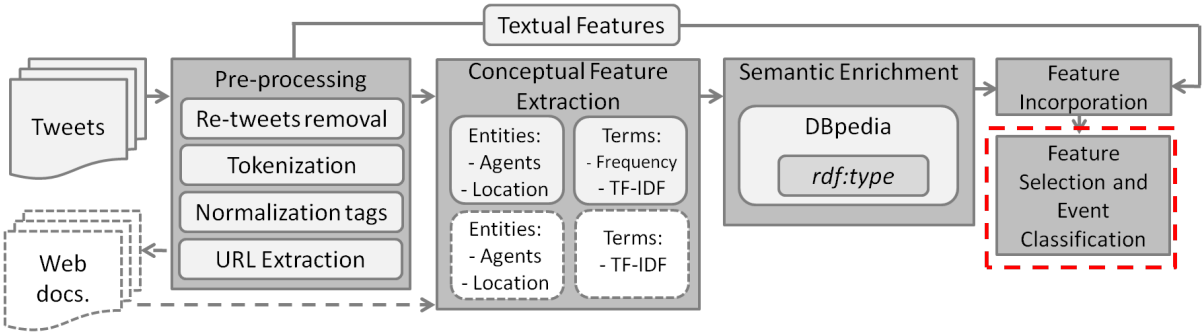
Each one of these datasets was prepared according to variations of the proposed framework, to evaluate the contribution of the key stages proposed, as follows:

Figure 5.2: Summarized pipeline of the Event Classification process, focusing on the Semantic Feature Pruning step



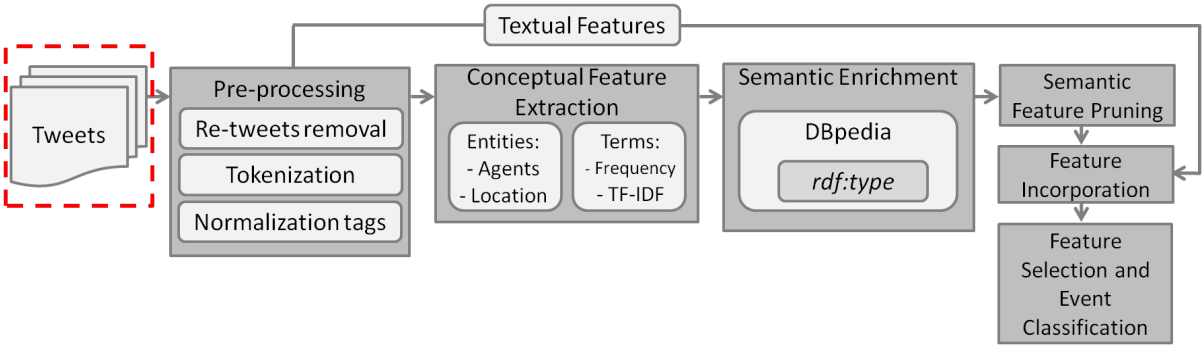
Source: the author.

Figure 5.3: Summarized pipeline of the Event Classification process, focusing on the Feature Selection step



Source: the author.

Figure 5.4: Summarized pipeline of the Event Classification process, without external document enrichment



Source: the author.

- 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, as presented in Figure 5.2.

- Experiment #1.2: evaluate the contribution of the Feature Selection step. To assess its complementary role to the Semantic Feature Pruning step, datasets are prepared according to all stages of the process (Figure 5.1), as well as all stages of the process except for

the Semantic Feature Pruning step, as described in Figure 5.3. We evaluate two feature selection algorithms, namely *CfsSubsetEval* and *InformationGain*;

- Experiment #1.3: given that discriminative features can be selected by the appropriate methods defined in the two above experiments, we evaluate the contribution of external documents enrichment in addition to semantic enrichment. Thus data is pre-processed according to all steps of the hybrid semantic enrichment framework (i.e. Figure 5.1), as well as without the extraction of the conceptual features from external web documents, as presented in Figure 5.4.

To classify the events, we used the algorithms available in Weka (HALL et al., 2009) (version 3.8.0), NB and SMO with PolyKernel, and a 10-fold cross-validation configuration. 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.

5.4 Dataset Preparation

We prepared the datasets according to the proposed framework (Figure 5.1), but according to the goal of the experiments, some of the steps were not executed. Table 5.2 summarizes this information.

Pre-processing: re-tweets were removed and normalization of specific features was performed to reduce sparsity (e.g. *@User* into *T_USER*). To perform the *external enrichment* the URLs mentioned in the tweets were recovered. We employed Text Extraction of AlchemyAPI[5] to extract the important information from this set of URLs, since this tool presented the best results in our analysis.

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 was also extracted the external web documents. We combined these conceptual features in three lists containing: a) only the named entities (i.e. agents and locations); b) only the frequent and representative terms, and c) the combination of both lists. To avoid duplicates, we removed: a) terms appearing in both frequent and representative lists and b) terms similar to named entities (e.g. for the frequent term drogba and the agent Didier Drogba, we kept just the entity).

[5]http://www.alchemyapi.com/products/alchemylanguage/text-extraction

We employed Open Calais API 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, 2016a). 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 (VOSECKY et al., 2014).

Semantic Enrichment: we used the DBpedia knowledge base and the *rdf:type* property. We employed the RapidMiner platform[6] (version 7.0), and the LOD Extension operators: a) DBpedia Spotlight to connect the conceptual features to the respective URI from DBpedia knowledge base and b) Direct Type operator.

Semantic Feature Pruning: we applied the PageRank-based pruning method to select the relevant semantic features, using the QUARTILES strategy as the method to automatically define the pruning thresholds, as described in Section 4.5 (Algorithm 3).

Incorporation: we combined the (pruned) semantic features with the textual features extracted from pre-processed tweets, according to the augmentation technique presented in Section 3.4.

Feature Selection: we applied a general-purpose feature selection technique, as an attempt to select the most discriminative textual and semantic features. We tested two techniques, namely *CfsSubsetEval* and *InformationGain*.

Table 5.2: 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	NA	Hybrid semantic enrichment	NA	Hybrid semantic enrichment	NA	Semantic-only, Hybrid semantic enrichment
Strategies for selecting features analyzed	NA	QUARTILES	CFS and InfoGain	CFS, InfoGain, QUARTILES+CFS, QUARTILES+IG	CFS	CFS, QUARTILES+CFS

Source: the author.

[6]http://rapidminer.com/

5.4.1 Experiment #1.1: The Semantic Feature Pruning Step

In this section, we report the performance of the proposed framework using the PageRank-based semantic feature pruning algorithm in combination to the QUARTILES strategy against the baseline. The QUARTILES strategy for the automatically definition of the thresholds was chosen due to the best results produced in our experiments, and for considering the characteristics and relationships of the semantic features, given the whole content of LOD cloud dataset used. Additional experiments using the IQR strategy for threshold definition are reported in Appendix B.

The datasets were prepared according to the process depicted in Figure 5.2, i.e. *hybrid semantic enrichment* configuration, pruning and no feature selection, as described in Table 5.2.

5.4.1.1 Results

In Table 5.3, we present the amount of textual features (*TF row*) for the baseline of each event type (i.e. without contextual enrichment), the amount of textual and semantic features resulting from the Incorporation step (*WP row*) without applying the pruning step, and the amount of textual and semantic features resulting from the Incorporation step after the application of PageRank-based pruning algorithm and the QUARTILES strategy (*QUARTILES row*). These results refer to the ALL variation of each dataset.

Table 5.3: 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

Source: the author.

Using the PageRank-based pruning algorithm and the QUARTILES strategy for defining the thresholds (from now on referred as pruning, for short), the average reduction in the number of features was about 41%. A qualitative analysis was not performed since the number of features was still elevated. So, we used the NB and SMO classification algorithms to compare the performance of using this kind of contextual enrichment and threshold against the baseline. Towards a complete analysis, we performed a statistical test, in which we claim that the improvement is *significant* with a significance level of $\alpha = 0.05$.

Table 5.4 presents the results for the *Positive* class, since we aim at identifying the tweets related to a specific event. The 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.

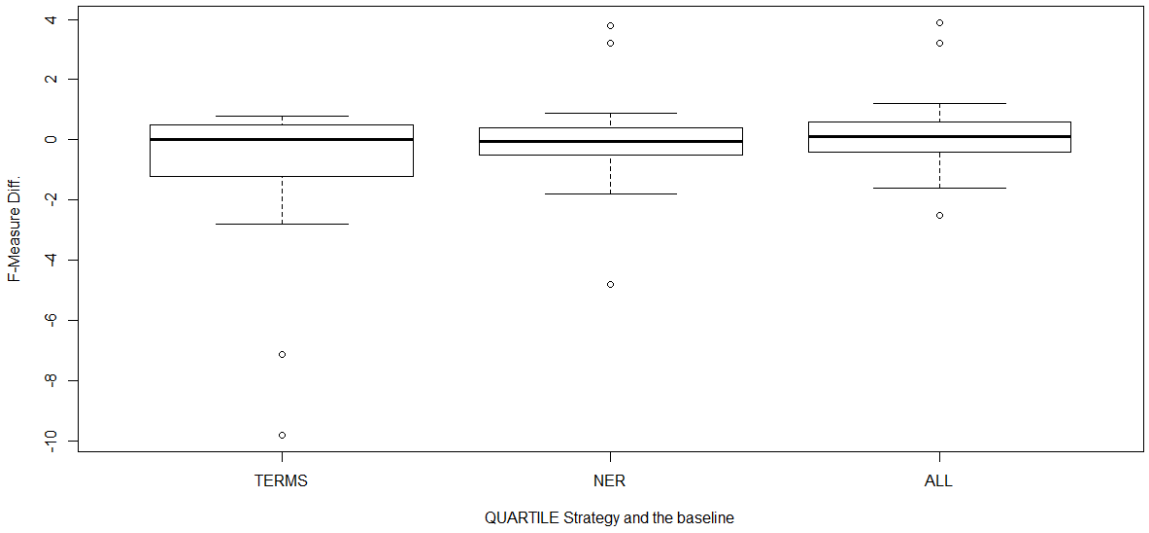
The boxplot of Figure 5.5 summarizes the results achieved for the F-measure, calculated by extracting the difference between the proposed approach and the baseline.

Table 5.4: Statistical comparison between the baseline and the hybrid semantic enrichment configuration with pruning only

Dataset	Algor.	Baseline			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.936	0.809	0.867	**0.979 v**	0.634 *	0.769 *	**0.961 v**	0.763 *	0.850	**0.961 v**	0.783	0.862
	SMO	0.940	0.909	0.924	0.946	0.912	0.929	0.944	0.913	0.928	0.947	0.913	0.930
Olympics	NB	0.724	0.713	0.717	**0.775 v**	0.556 *	0.646 *	0.678 *	**0.782 v**	0.726	0.682 *	**0.784 v**	0.729
	SMO	0.885	0.823	0.853	0.881	0.823	0.850	0.881	0.819	0.848	0.882	0.819	0.849
Halloween	NB	0.859	0.733	0.790	0.844	0.723	0.778	0.835	0.719	0.772	0.832	0.725	0.774
	SMO	0.896	0.888	0.892	0.893	0.886	0.889	0.895	0.889	0.891	0.895	0.889	0.891
HSandy	NB	0.917	0.848	0.881	**0.949 v**	0.776 *	0.853 *	0.873 *	0.798 *	0.833 *	0.884 *	0.830	0.856 *
	SMO	0.966	0.919	0.942	0.969	0.923	0.946	0.962	0.919	0.940	0.961	0.921	0.940
Alberta Floods	NB	0.945	0.952	0.948	**0.970 v**	0.938	0.953	**0.979 v**	**0.981 v**	**0.980 v**	**0.978 v**	**0.981 v**	**0.980 v**
	SMO	0.999	0.992	0.995	0.999	0.991	0.995	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
Australia Bushfire	NB	0.931	0.960	0.945	0.948	0.960	0.953	**0.979 v**	**0.989 v**	**0.983 v**	**0.980 v**	**0.988 v**	**0.984 v**
	SMO	0.999	0.992	0.995	0.997	0.993	0.995	0.998	0.998	0.998	0.997	0.996	0.997
Influenza	NB	0.961	0.998	0.979	**0.977 v**	0.997	**0.987 v**	0.955	0.997	0.976	**0.973 v**	0.995	0.984
	SMO	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999

Source: the author.

Figure 5.5: Difference between hybrid semantic enrichment with pruning only and the baseline, considering the F-Measure metric



Source: the author.

5.4.1.2 Discussions

Regarding the baseline, it is important to stress that SMO produces better results, compared to NB. The performance for some cases are quite good, and therefore sometimes there is room only for marginal improvements.

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 42% of the cases. An improvement is given by the difference between the results achieved with our approach and the baseline, considering the same metric. For example, considering the TERMS variation and the F-Measure metric for the Australia Bushfire dataset, our approach yielded a better result (0.953) compared to the baseline (0.945), with an improvement of 0.5 percentage points (pp). In general, improvements range from 0.1 pp to 7.1 pp in specific cases (i.e. the Olympics dataset, considering the Recall metric and the ALL type of feature).

However, we observed that our approach statistically outperformed the baseline only in 21.4% of the cases. The NB algorithm, the Alberta Floods target event, and the Precision metric were the cases in which more statistically significant improvements were noticed. Our previous work has already shown that the performance of the NB algorithm was more affected by the number of features (ROMERO; BECKER, 2016a).

According to the boxplot in Figure 5.5, the Semantic Feature Pruning step produced more improvements in the NER and ALL variations, since they present a lower dispersion, upper quartiles near 0.6 pp, and some outliers around 4 pp. The median improvement for the ALL variation is 0.1 pp.

In summary, the proposed hybrid approach using pruning as the only means to reduce the number of semantic features produced marginal improvements in the event classification performance.

5.4.2 Experiment #1.2: the Feature Selection Step

The proposed framework with the Semantic Feature Pruning step as the only means to select discriminative features was able to improve the results, but only marginally. In this section we perform experiments to evaluate whether it could be replaced by a feature selection technique, or be used in combination with one. We claim that they address complementary issues: whereas pruning selects the most representative semantic features from the Semantic Enrichment step 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 prepared our variations for each TERMS, NER and ALL dataset, using the processes depicted in Figures 5.3 and 5.1 (Table 5.2):

- CFS: without pruning, using only the *CfsSubsetEval* feature selection algorithm and BestFirst as the search method;
- InfoGain: without pruning, using only the *InformationGain* feature selection algorithm, using 0 as threshold;
- QUARTILES+CFS: combination of the PageRank-based algorithm according to the QUARTILES strategy and the *CfsSubsetEval* feature selection algorithm and BestFirst as the search method;
- QUARTILES+IG: combination of the PageRank-based algorithm according to the QUARTILES strategy and the *InformationGain* feature selection algorithm, using 0 as threshold.

As baseline, we prepared variations of the original baseline datasets by applying the *CfsSubsetEval* and the *InformationGain* feature selection algorithms.

5.4.2.1 Results

Table 5.5 presents, for each dataset, the number of textual/semantic features without applying techniques for selecting the relevant features (row *WP*), the number of textual and semantic features resulting from the application of each feature selection algorithm (rows *InfoGain* and *CFS*), as well as their combination with the pruning algorithm (rows *QUARTILES+IG* and *QUARTILES+CFS*). These results refer to the ALL variation.

We then applied the NB and SMO algorithms for all these variations, and analyzed the performance of each variation against the baseline using two-tail paired $t-test$ with significance level $\alpha = 0.05$. In this way, we can analyze the contribution of the proposed *hybrid semantic enrichment* with regard to the influence of the feature selection algorithm. Tables 5.6, 5.7, 5.8, and 5.9, summarize the results for the *Positive* class.

The boxplots of Figures 5.6 and 5.7 summarize the results achieved for the F-Measure, calculated by extracting the difference between the compared approaches (i.e. feature selection only, pruning combined with feature selection) and the baseline.

Table 5.5: Number of features resulting from the application of the different techniques to select the most relevant features, considering the ALL variation

Dataset	FaCup	Olympics	Halloween	HSandy	Alberta F.	Australia B.	Influenza
WP	2182	3723	4197	4311	5068	4055	2657
InfoGain	1242	1471	1717	1891	1895	1616	1204
CFS	77	88	131	71	29	46	52
QUARTILES+IG	943	880	1076	1234	1030	950	1021
QUARTILES+CFS	70	101	146	70	29	58	53

Source: the author.

Table 5.6: Statistical comparison between the baseline and the hybrid semantic enrichment, both using only the*CfsSubsetEval* algorithm

Dataset	Algor.	Baseline+CFS			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.967	0.730	0.832	0.947 *	**0.770 v**	0.849	0.930 *	**0.789 v**	0.853	0.941 *	0.707	0.806
	SMO	0.978	0.769	0.861	0.962 *	**0.826 v**	**0.889 v**	0.969	**0.819 v**	**0.887 v**	0.966	**0.822 v**	**0.888 v**
Olympic	NB	0.970	0.454	0.616	0.786 *	0.476	0.592	0.884 *	**0.541 v**	**0.670 v**	0.826 *	**0.520 v**	0.637
	SMO	0.956	0.610	0.744	0.930	0.620	0.743	0.936	**0.654 v**	0.770	0.938	**0.663 v**	**0.776 v**
Halloween	NB	0.805	0.827	0.816	0.840	0.801	0.819	0.803	**0.867 v**	0.833	0.808	**0.867 v**	0.836
	SMO	0.840	0.877	0.858	0.832	0.857	0.844	0.847	0.874	0.860	0.846	0.879	0.862
HSandy	NB	0.976	0.822	0.892	0.983	0.823	0.896	0.922 *	0.843	0.881	0.912 *	**0.854 v**	0.882
	SMO	0.977	0.873	0.922	0.984	0.872	0.924	0.970	0.832 *	0.896 *	0.957 *	0.861	0.906
Alberta Flood	NB	0.991	0.971	0.981	0.982	**0.991 v**	0.987	0.996	**1.000 v**	**0.998 v**	0.996	**1.000 v**	**0.998 v**
	SMO	0.999	0.987	0.993	0.990 *	0.987	0.988	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
Australia Bushfire	NB	0.957	0.944	0.950	0.925 *	**0.979 v**	0.951	**0.982 v**	**0.993 v**	**0.987 v**	**0.982 v**	**0.994 v**	**0.988 v**
	SMO	0.992	0.981	0.986	0.979	0.982	0.980	0.992	**0.997 v**	**0.994 v**	0.992	**0.996 v**	0.994
Influenza	NB	0.999	0.997	0.998	0.999	0.998	0.998	0.999	0.996	0.997	0.999	0.995	0.997
	SMO	1.000	0.997	0.998	1.000	0.997	0.999	1.000	0.997	0.998	1.000	0.997	0.998

Source: the author.

Table 5.7: Statistical comparison between the baseline and the hybrid semantic enrichment, both using the *InformationGain* algorithm

Dataset	Algor.	Baseline+IG			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.937	0.814	0.871	0.936	0.587 *	0.720 *	0.937	0.658 *	0.773 *	0.910 *	0.751 *	0.823 *
	SMO	0.972	0.910	0.940	0.975	0.908	0.940	0.971	0.914	0.941	0.970	0.912	0.940
Olympic	NB	0.769	0.691	0.727	0.792	0.244 *	0.372 *	0.498 *	**0.756 v**	0.600 *	0.592 *	0.611 *	0.601 *
	SMO	0.942	0.832	0.883	0.930	0.832	0.878	0.928	0.845	0.883	0.923	0.845	0.881
Halloween	NB	0.859	0.754	0.803	0.847	0.746	0.793	0.582 *	0.705 *	0.637 *	0.605 *	0.705 *	0.651 *
	SMO	0.921	0.904	0.912	0.921	0.908	0.914	0.917	0.904	0.910	0.913	0.905	0.909
HSandy	NB	0.927	0.854	0.889	0.867 *	0.667 *	0.753 *	0.756 *	0.769 *	0.762 *	0.793 *	0.794 *	0.793 *
	SMO	0.977	0.923	0.949	0.976	0.933	0.954	0.969	0.925	0.946	0.968	0.927	0.947
Alberta Flood	NB	0.955	0.954	0.954	0.994	0.546 *	0.690 *	**0.992 v**	**0.985 v**	**0.989 v**	**0.992 v**	**0.985 v**	**0.989 v**
	SMO	1.000	0.998	0.999	0.998	0.995	0.997	0.998	1.000	0.999	0.998	1.000	0.999
Australia Bushfire	NB	0.937	0.960	0.948	0.920	0.654 *	0.763 *	0.955	**0.989 v**	**0.972 v**	0.951	**0.993 v**	**0.971 v**
	SMO	0.996	0.995	0.996	0.997	0.994	0.996	0.996	0.996	0.996	0.997	0.997	0.997
Influenza	NB	0.967	0.998	0.982	**0.995 v**	0.779 *	0.873 *	0.835 *	0.939 *	0.884 *	0.888 *	0.911 *	0.899 *
	SMO	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999

Source: the author.

Table 5.8: Statistical comparison between the baseline and the hybrid semantic enrichment combining pruning and the *CfsSubsetEval* algorithms

Dataset	Algor.	Baseline+CFS			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.967	0.730	0.832	0.953	**0.790 v**	**0.863 v**	0.955	0.747	0.838	0.926 *	**0.803 v**	**0.860 v**
	SMO	0.978	0.769	0.861	0.970	**0.811 v**	**0.883 v**	0.968	**0.806 v**	0.879	0.968	**0.815 v**	**0.885 v**
Olympics	NB	0.970	0.454	0.616	0.876 *	0.414	0.561 *	0.664 *	**0.780 v**	**0.703 v**	0.664 *	**0.780 v**	**0.703 v**
	SMO	0.956	0.610	0.744	0.921 *	0.630	0.747	0.951	0.621	0.751	0.951	0.621	0.751
Halloween	NB	0.805	0.827	0.816	0.802	0.827	0.814	0.809	0.847	0.827	0.809	0.847	0.827
	SMO	0.840	0.877	0.858	0.842	0.867	0.854	0.847	0.874	0.860	0.847	0.874	0.860
HSandy	NB	0.976	0.822	0.892	0.886 *	**0.886 v**	0.886	0.903 *	0.829	0.864 *	0.908 *	0.848	0.877
	SMO	0.977	0.873	0.922	0.918 *	**0.906 v**	0.912	0.955 *	0.841 *	0.894 *	0.952 *	0.849	0.898 *
Alberta Flood	NB	0.991	0.971	0.981	0.991	0.973	0.982	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
	SMO	0.999	0.987	0.993	0.999	0.989	0.994	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
Australia Bushfire	NB	0.957	0.944	0.950	0.951	**0.982 v**	0.966	**0.980 v**	**0.987 v**	**0.983 v**	**0.980 v**	**0.987 v**	**0.983 v**
	SMO	0.992	0.981	0.986	0.990	0.979	0.984	0.981	**0.998 v**	0.989	0.985	**0.999 v**	0.992
Influenza	NB	0.999	0.997	0.998	0.999	0.998	0.998	0.999	0.997	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	1.000	0.997	0.999

Source: the author.

Table 5.9: Statistical comparison between the baseline and the hybrid semantic enrichment combining pruning and the *InformationGain* algorithms

Dataset	Algor.	Baseline+IG			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.937	0.814	0.871	**0.979 v**	0.631 *	0.767 *	**0.962 v**	0.768 *	0.853	**0.960 v**	0.786	0.864
	SMO	0.972	0.910	0.940	0.974	0.910	0.941	0.971	0.909	0.939	0.971	0.912	0.940
Olympics	NB	0.769	0.691	0.727	0.796	0.534 *	0.638 *	0.704 *	**0.771 v**	0.735	0.704 *	**0.768 v**	0.733
	SMO	0.942	0.832	0.883	0.933	0.831	0.879	0.938	0.838	0.885	0.936	0.840	0.885
Halloween	NB	0.859	0.754	0.803	0.847	0.742	0.790	0.840	0.733	0.782	0.838	0.736	0.783
	SMO	0.921	0.904	0.912	0.919	0.905	0.912	0.923	0.906	0.914	0.922	0.907	0.914
HSandy	NB	0.927	0.854	0.889	**0.961 v**	0.778 *	0.860 *	0.873 *	0.801 *	0.836 *	0.884 *	0.832	0.857 *
	SMO	0.977	0.923	0.949	0.974	0.932	0.952	0.976	0.924	0.949	0.976	0.929	0.951
Alberta Flood	NB	0.955	0.954	0.954	**0.974 v**	0.945	0.959	**0.979 v**	**0.984 v**	**0.982 v**	**0.979 v**	**0.984 v**	**0.982 v**
	SMO	1.000	0.998	0.999	0.999	0.997	0.998	0.998	1.000	0.999	0.998	1.000	0.999
Australia Bushfire	NB	0.937	0.960	0.948	0.956	0.964	0.960	**0.979 v**	**0.989 v**	**0.983 v**	**0.980 v**	**0.988 v**	**0.984 v**
	SMO	0.996	0.995	0.996	0.998	0.996	0.997	0.998	0.998	0.998	0.997	0.997	0.997
Influenza	NB	0.967	0.998	0.982	**0.980 v**	0.997	0.989	0.963	0.998	0.980	0.977	0.995	0.985
	SMO	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999

Source: the author.

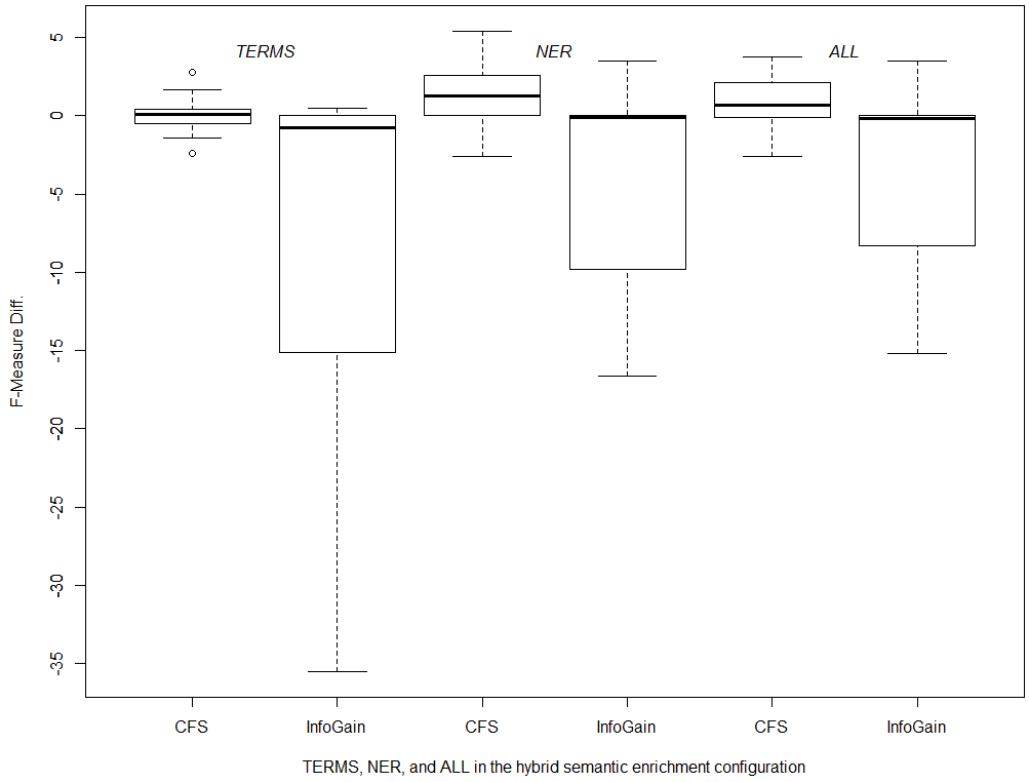
5.4.2.2 Discussions

a) Quantitative and Qualitative Analysis of the Different Feature Selection Techniques

In Table 5.5, we can observe the impact of each feature selection algorithm in the set of semantic and textual features. Recall that these features can be semantic, and thus resulting from the enrichment of positive examples, or textual, i.e. tokens extracted from all tweets, even the ones not related to the target event. We can observe that the *CfsSubsetEval* selects a significantly smaller number of features compared to *InformationGain*. Indeed, the former considers additionally the redundancy among features as a selection criteria. Note that these results may be influenced by the choice of parameters.

With regard to the contribution of pruning in this reduction, we observe that the number of features for QUARTILES+CFS is very similar to the application of the *CfsSubsetEval* algorithm only. The reduction is substantially larger in the case of the *InformationGain* algorithm, as the results of using only this algorithm are, in average, 27.5% bigger than its combination with pruning.

Figure 5.6: Difference of performance between Hybrid Semantic Enrichment using *CfsSubsetEval/InformationGain* algorithms and the baseline, considering the F-Measure metric



Source: the author.

The resulting list of textual/semantic features is organized according to the feature selection algorithm criterion, such that the most representative ones are in the top of the list. We manually analyzed the content of these lists to observe the characteristics of the resulting features.

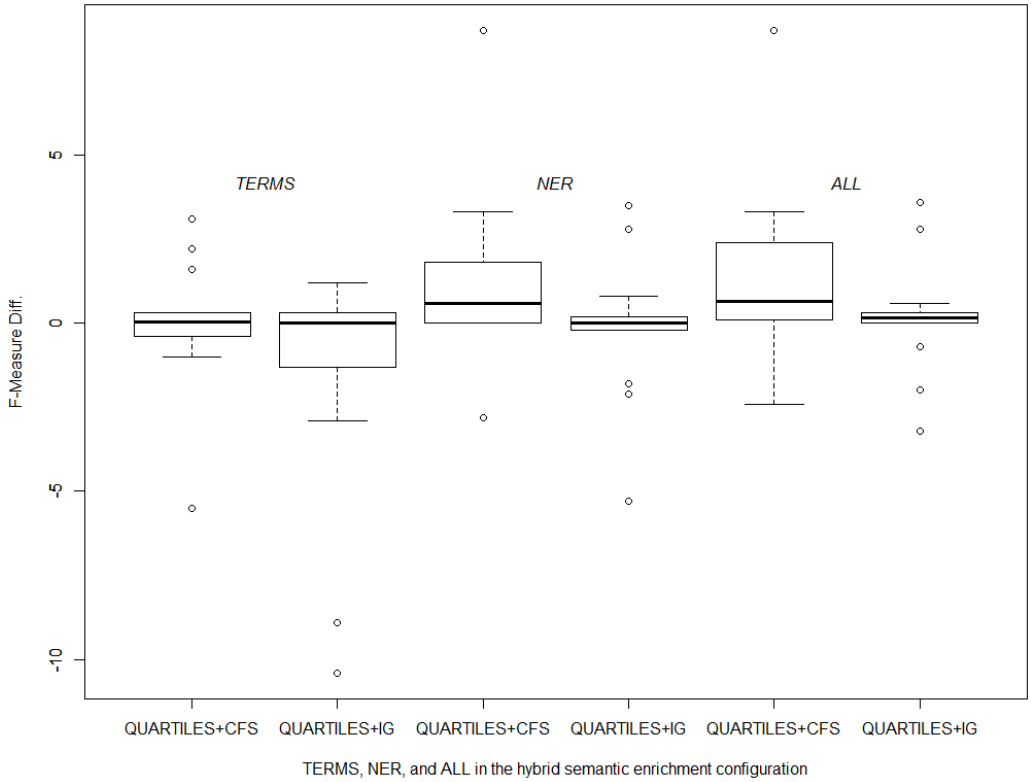
Considering the application of the feature selection algorithms only, we observed that for the FaCup dataset the semantic and textual features appearing in the top of the list produced by the *InformationGain* algorithm are very representative of the domain (e.g. *ontology/SportsTeam*, #cfc, #lfc, chelsea). Considering the *CfsSubsetEval* algorithm, just a few semantic features were selected, but they are representative of the domain analyzed.

In the Olympics dataset, a mixed list of textual and semantic features belonging to positive and negative examples was produced for both algorithms. The same pattern could be noticed in the results of the natural disaster datasets HSandy and Alberta Floods.

Analyzing the Halloween dataset, we observed that most of the resulting features, for both algorithms, are very related to the domain of the event analyzed. The same pattern could be noticed in the results of the natural disaster and epidemics datasets, Australia Bushfire and Influenza, respectively.

In summary, both feature selection algorithms were able to select relevant textual and

Figure 5.7: Difference of performance between Hybrid Semantic Enrichment using Pruning and *CfsSubsetEval*/*InformationGain* algorithms and the baseline, considering the F-Measure metric



Source: the author.

semantic features, but the *CfsSubsetEval* algorithm presented the highly related features to the event target in the top of the resulting list. Regarding the pruning algorithm, we observed that it selects the semantic features that better generalizes the event domain, such that its combination 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) Feature Selection Algorithms

Tables 5.6 and 5.7 detail the results of the hybrid semantic enrichment approach considering the application of a general-purpose feature selection algorithm only (i.e. without the pruning algorithm). Recall that we applied the respective feature selection algorithm to the baseline for this comparison.

Regarding the feature selection algorithm, we noticed that the *CfsSubsetEval* presented the best results in comparison to the use of the *InformationGain* algorithm. With its adoption, we outperformed the baseline in 52.3% of the cases, producing improvements that range from 0.1 pp to 8.7 pp in specific situations (i.e. the NER variation and the Recall metric). However, these improvements were statistically significant in only 28.5% of the cases, mainly for the

Recall metric and the Alberta Floods and Australia Bushfire datasets.

The results using *InformationGain* are substantially inferior (Table 5.7). In this case improved results were observed in 28.5% of the cases, ranging from 0.1 pp to 6.5 pp in specific situations (i.e. the NER variation and the Recall metric). Considering the statistical analysis, our approach outperformed the baseline in almost 10% of the cases, mainly for the Alberta Floods and Australia Bushfire datasets.

In Figure 5.6, we present the boxplot comparing both feature selection algorithms, considering the F-Measure metric. The boxplot was built considering the difference between the enriched dataset and the baseline, according to the respective feature selection algorithm. We can observe that the combination of the *CfsSubsetEval* feature selection algorithm with *hybrid semantic enrichment* produced good results, with a median improvement of 1.25 pp and a maximum value of 5.4 pp. The best results were achieved with the NER variation. We can also notice that, in this situation, the utilization of the TERMS variation produced a slightly decrease in the results. The combination with the *InformationGain* algorithm produced results with a high dispersion and a negative distribution.

Considering the Semantic Feature Pruning and the Feature Selection steps in isolation, these results demonstrated that using only the general-purpose feature selection algorithm (i.e. the *CfsSubsetEval*) to select the relevant semantic and textual features of the dataset, produced slightly better results than using only the pruning algorithm as the only means to select the discriminatory features.

c) Combination of Semantic Feature Pruning and Feature Selection

Finally, we consider the contribution of performing both the Semantic Feature Pruning and the Feature Selection steps on the classification peformance. As presented in Table 5.8, the combination of PageRank-based pruning algorithm and the QUARTILES strategy with the *CfsSubsetEval* algorithm was able to statistically outperform the baseline in 25.4% of the cases. In general, the improvements could be noticed in 53.17% of the cases, most of them for the Recall metric and in the FaCup, Alberta Floods and Australia Bushfire datasets. Considering all the results, the average improvement was about 3 pp, ranging from 0.1 pp to 32.6 pp in specific cases (i.e. for the Olympics dataset, considering the ALL variation, and the Recall metric).

When considering the combination of *InformationGain* and pruning, we were able to statistically outperform the baseline in 15.8% of the cases, as described in Table 5.9. In this situation the maximum improvement was 7.7 pp for the Olympics target event, using all types of features proposed (i.e. the ALL variation). Most of the improvements could be noticed for

the Precision metric. In general, improvements could be noticed in almost 47.6% of the cases. Thus, the combination of the pruning algorithm with this particular feature selection algorithm significantly improved the results.

In Figure 5.7, we can observe the boxplots of each combination, considering the F-Measure metric. In both strategies, the best results were achieved with the ALL variation, with a maximum value of 8.7 pp and 3.6 pp for the QUARTILES+CFS and QUARTILES+IG combinations, respectively. The upper quartile of the TERMS variation were the same in both combinations. For the QUARTILES+CFS combination, the upper quartile of the NER and ALL type of features resulted in 1.62 pp and 2.25 pp, respectively. For the QUARTILES+IG combination, these results were 0.25 pp and 0.2 pp, respectively.

d) Final Comparative Analysis

In Table 5.10, we summarize all these analyses through the presentation of the results for the *hybrid semantic enrichment* without the application of any pruning technique or feature selection algorithm (*Without pruning* row), with the application of the *CfsSubsetEval* algorithm only (*CFS only* row), with the application of the *InformationGain* algorithm only (*InfoGain only* row), with the combination of the pruning with the *CfsSubsetEval* algorithm (*QUARTILES+CFS* row), and with the combination of the pruning with the *InformationGain* algorithm (*QUARTILES+IG* row). For each variation, we analyzed the general improvement, the percentage of statistically superior results, the percentage of results that was statistically outperformed by the baseline, the minimum improvement achieved, and the maximum improvement achieved, the latter two in percentage points. Considering the statistical analysis, we also highlight the most improved type of feature (i.e. TERMS, NER, and ALL), the most improved metric, and the most improved target events[7].

The results using only the *CfsSubsetEval* feature selection algorithm are very similar to its combination with the pruning algorithm. The latter produced slightly better results considering the general improvements, and a significant better maximum improvement was achieved. Considering the distributions in the boxplots for both approaches, we can observe that despite the medians are very similar, the lower quartiles of the QUARTILES+CFS combination are

[7]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 both classification algorithms. For example, in the QUARTILES+CFS combination, the FaCup and Alberta Flood datasets statistically outperformed the baseline in nine cases, each one. Thus, to be inserted in this table, the others datasets must have, at least, 4.5 cases that statistically outperformed the baseline, considering both classification algorithms.

Table 5.10: Summarization of all results

Configuration	General improvement	Statistically superior	Statistically inferior	Minimum improvement	Maximum improvement	Most improved type of feature	Most improved metric	Most improved dataset
Without Pruning	31.0%	12.7%	31.0%	0.1 pp	6.9 pp	NER and ALL	Recall and F-Measure	Alberta Floods
CFS only	52.3%	28.5%	11.0%	0.1 pp	8.7 pp	NER	Recall	FaCup, Olympics, Australia B., Alberta F.
InfoGain only	28.5%	10.0%	32.5%	0.1 pp	6.5 pp	NER	Recall	Alberta Floods
QUARTILES only	42.0%	21.4%	11.0%	0.1 pp	7.1 pp	ALL	Precision	Australia B. Alberta F.
QUARTILES+CFS	53.1%	25.4%	12.7%	0.1 pp	32.6 pp	ALL	Recall	FaCup, Australia B., Alberta F.
QUARTILES+IG	47.6%	15.8%	11.9%	0.1 pp	7.7 pp	NER and ALL	Precision	Australia B., Alberta F.

Source: the author.

closer to zero (0), compared to the ones of *CfsSubsetEval* algorithm, which means that the latter produced more losses. The ALL variation produced the best results considering mainly the upper quartiles and the maximum value achieved by the QUARTILES+CFS combination (i.e. 2.2 pp and 8.7 pp, respectively, against 2 pp and 3.8 pp from using only the *CfsSubsetEval* algorithm).

In all the feature types combinations, the presence of the named entities produced the best results (i.e. NER and ALL). Regarding the metrics, in almost all cases the largest number of improvements could be noticed for the Recall metric, thus, in general, the hybrid semantic approach with properly selected features presented good performance in recognizing the tweets related to the target event.

5.4.3 Experiment #1.3: Semantic-only Enrichment vs. Hybrid Semantic Enrichment

After performing the analysis of the different configurations for selecting the most relevant textual and semantic features, we performed a final comparison between the *hybrid semantic enrichment* and the *semantic-only enrichment* configurations. In the former, we employ the conceptual features extracted from the tweet text and the related web documents, recovered through the URLs mentioned in the tweets, according to the process of Figure 5.1. In the latter, the conceptual features are extracted only from the tweet text to be used as input to the Semantic Enrichment step, as depicted in Figure 5.4. Table 5.2 summarizes the preparation steps for these datasets. Given the close results reported in the previous sections for using the *CfsSubsetEval* algorithm only and its combination with pruning (i.e. QUARTILES+CFS), we decided to develop the evaluation of these two configurations.

5.4.3.1 Results

In Table 5.11, for each type of feature, we present the amount of textual and semantic features resulting from the Incorporation step (*Without pruning* column), the amount of semantic and textual features resulting from the application of the *CfsSubsetEval* algorithm (*CFS* column), and the amount of semantic and textual features resulting from the Semantic Feature Pruning and Feature Selection steps (*QUARTILE+CFS* column). This analysis was performed for the *semantic-only enrichment* (*SOE* rows) and the *hybrid semantic* enrichment (*HSE* rows).

We then applied the NB and SMO classification algorithms for each dataset in both enrichment approaches, and performed a statistical test to evaluate the results. The results for the *semantic-only enrichment* are presented in Tables 5.12 and 5.13, considering the results for the *Positive* class. The respective results for the *hybrid semantic enrichment* are available in Tables 5.6 and 5.8. Notice that the baseline was produced by employing the *CfsSubsetEval* technique, since all these enrichment datasets were produced using the same algorithm.

In Figures 5.8 and 5.9, we present boxplots that compare these enrichment configurations, considering both methods for selecting discriminative features (i.e. CFS and QUARTILES+CFS). They were produced by calculating the difference between the enrichment approach with the corresponding technique for selecting discriminative features and the baseline.

Finally, we performed an analysis to evaluate the boost produced by using external documents in the enrichment step in comparison to the *semantic-only enrichment* approach. In other words, we considered the *semantic-only enrichment* as the baseline, to analyze the contribution of the *hybrid semantic enrichment* configuration for each target event.

Figures 5.10, 5.11, 5.12 and 5.13, present the average difference between the *hybrid semantic* and *semantic-only* enrichment in percentage points, considering the NB and SMO classification algorithms, and the Precision, Recall, and F-Measure metrics. These results refer to the application of the *CfsSubsetEval* algorithm only (Figures 5.10 and 5.11), and its combination to pruning algorithm (Figures 5.12 and 5.13).

Table 5.11: Amount of textual and semantic features for each configuration

Dataset	Contextual Enrich.	Without pruning			CFS			QUARTILES+CFS		
		TERMS	NER	ALL	TERMS	NER	ALL	TERMS	NER	ALL
FaCup	SOE	1844	2036	2100	68	64	64	58	64	61
	HSE	1844	2134	2182	68	68	77	58	68	70
Olympics	SOE	1967	2381	2417	98	89	90	95	97	95
	HSE	1963	3703	3723	98	90	88	95	101	101
Halloween	SOE	1873	2974	2983	139	134	131	150	139	139
	HSE	1884	4180	4197	146	131	131	150	146	146
HSandy	SOE	2256	2526	2577	59	75	62	66	77	70
	HSE	2282	4286	4311	79	69	71	83	71	70
Alberta Floods	SOE	2066	2527	2552	52	52	50	57	52	52
	HSE	2108	5048	5068	56	29	29	53	29	29
Australia Bushfire	SOE	2251	2744	2785	62	48	44	55	48	47
	HSE	2346	3984	4055	58	46	46	54	53	58
Influenza	SOE	1981	2302	2371	53	46	53	55	46	55
	HSE	1981	2604	2657	53	44	52	53	46	53

Source: the author.

Table 5.12: Statistical comparison between the baseline and the semantic-only enrichment configuration, both using the only *CfsSubsetEval* algorithm

Dataset	Algor.	Baseline+CFS			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.967	0.730	0.832	0.947 *	**0.770 v**	0.849	0.940 *	0.761	0.841	0.918 *	**0.792 v**	0.850
	SMO	0.978	0.769	0.861	0.962 *	**0.826 v**	**0.889 v**	0.973	0.787	0.870	0.952 *	**0.822 v**	0.882
Olympics	NB	0.970	0.454	0.616	0.786 *	0.476	0.592	0.887 *	0.495	0.634	0.721 *	**0.543 v**	0.619
	SMO	0.956	0.610	0.744	0.930	0.620	0.743	0.946	0.639	0.762	0.922 *	0.635	0.751
Halloween	NB	0.805	0.827	0.816	0.815	0.832	0.823	0.820	0.831	0.825	0.826	0.835	0.830
	SMO	0.840	0.877	0.858	0.843	0.882	0.862	0.840	0.877	0.858	0.838	0.880	0.858
HSandy	NB	0.976	0.822	0.892	0.884 *	**0.928 v**	0.905	0.903 *	**0.871 v**	0.886	**0.988 v**	**0.883 v**	**0.932 v**
	SMO	0.977	0.873	0.922	0.998	**0.913 v**	**0.949 v**	**0.991 v**	0.838 *	0.908	0.986	**0.921 v**	**0.952 v**
Alberta Floods	NB	0.991	0.971	0.981	0.946 *	**0.997 v**	0.971	0.883 *	**1.000 v**	0.937 *	0.882 *	**1.000 v**	0.937 *
	SMO	0.999	0.987	0.993	0.990 *	0.987	0.989	1.000	0.989	0.994	1.000	0.989	0.994
Australia Bushfire	NB	0.957	0.944	0.950	**0.984 v**	**0.980 v**	**0.982 v**	0.899 *	**0.998 v**	0.946	0.897 *	**0.998 v**	0.945
	SMO	0.992	0.981	0.986	0.997	0.990	0.993	0.995	**0.995 v**	**0.995 v**	0.998	**0.997 v**	**0.998 v**
Influenza	NB	0.999	0.997	0.998	0.999	0.998	0.998	0.999	0.997	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	1.000	0.997	0.999

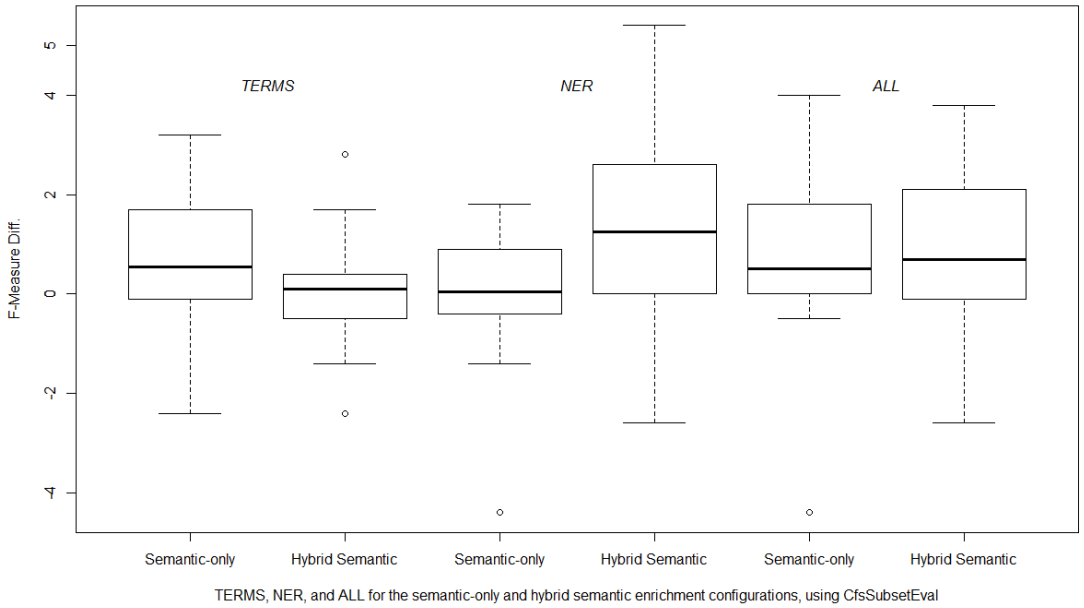
Source: the author.

Table 5.13: Statistical comparison between the baseline and the semantic-only enrichment configuration using the pruning algorithm in combination with the *CfsSubsetEval* algorithm

Dataset	Algort.	Baseline+CFS			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.967	0.730	0.832	0.953	**0.790 v**	**0.863 v**	0.940 *	0.761	0.841	0.970	0.762	0.853
	SMO	0.978	0.769	0.861	0.970	**0.811 v**	**0.883 v**	0.973	0.787	0.870	0.973	**0.801 v**	0.878
Olympics	NB	0.970	0.454	0.616	0.876 *	0.414	0.561 *	0.876 *	0.487	0.624	0.745 *	**0.526 v**	0.615
	SMO	0.956	0.610	0.744	0.921 *	0.630	0.747	0.945	0.626	0.753	0.943	0.613	0.742
Halloween	NB	0.805	0.827	0.816	0.802	0.827	0.814	0.801	0.820	0.810	0.801	0.820	0.810
	SMO	0.840	0.877	0.858	0.842	0.867	0.854	0.842	0.877	0.859	0.842	0.877	0.859
HSandy	NB	0.976	0.822	0.892	0.945	**0.879 v**	0.910	0.913 *	**0.865 v**	0.888	0.916 *	**0.881 v**	0.898
	SMO	0.977	0.873	0.922	0.985	**0.916 v**	**0.949 v**	**0.991 v**	0.854	0.917	0.936 *	**0.918 v**	0.927
Alberta Floods	NB	0.991	0.971	0.981	0.986	**0.985 v**	0.985	0.989	0.973	0.981	0.989	0.973	0.981
	SMO	0.999	0.987	0.993	0.993	0.989	0.991	0.999	0.988	0.993	0.999	0.988	0.993
Australia Bushfire	NB	0.957	0.944	0.950	0.938	**0.981 v**	0.959	0.881 *	**0.987 v**	0.931	0.841 *	**0.999 v**	0.913 *
	SMO	0.992	0.981	0.986	0.990	0.978	0.984	0.996	0.986	0.991	0.997	0.989	0.993
Influenza	NB	0.999	0.997	0.998	0.999	0.998	0.998	0.999	0.997	0.998	0.999	0.998	0.998
	SMO	1.000	0.997	0.998	1.000	0.997	0.998	1.000	0.997	0.998	1.000	0.997	0.998

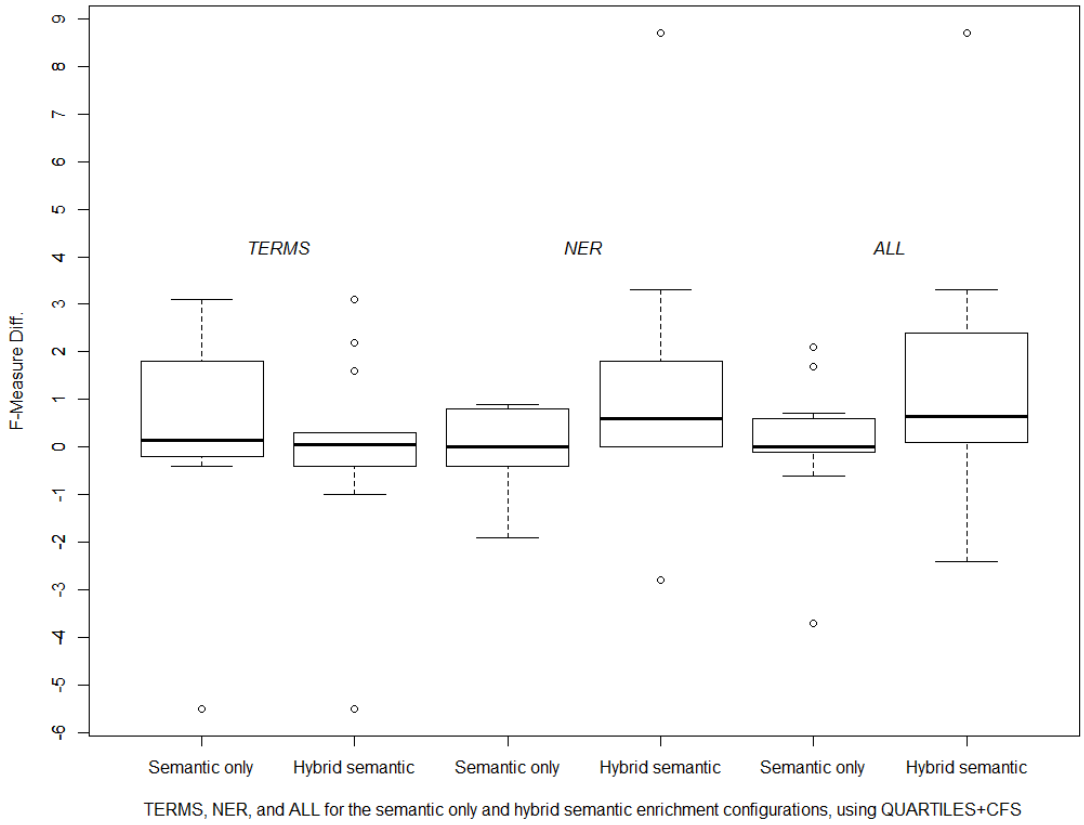
Source: the author.

Figure 5.8: Comparison of performance between the semantic-only and hybrid semantic enrichment using *CfsSubsetEval* algorithm, considering the F-Measure metric



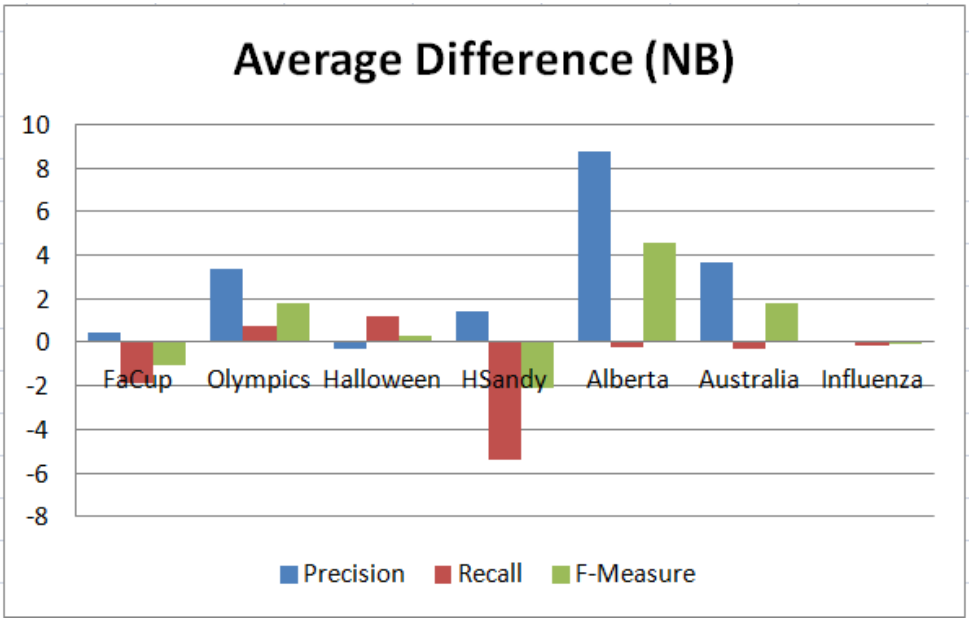
Source: the author.

Figure 5.9: Comparison of performance between the semantic-only and hybrid semantic enrichment using the combination of pruning and *CfsSubsetEval* algorithms, considering the F-Measure metric



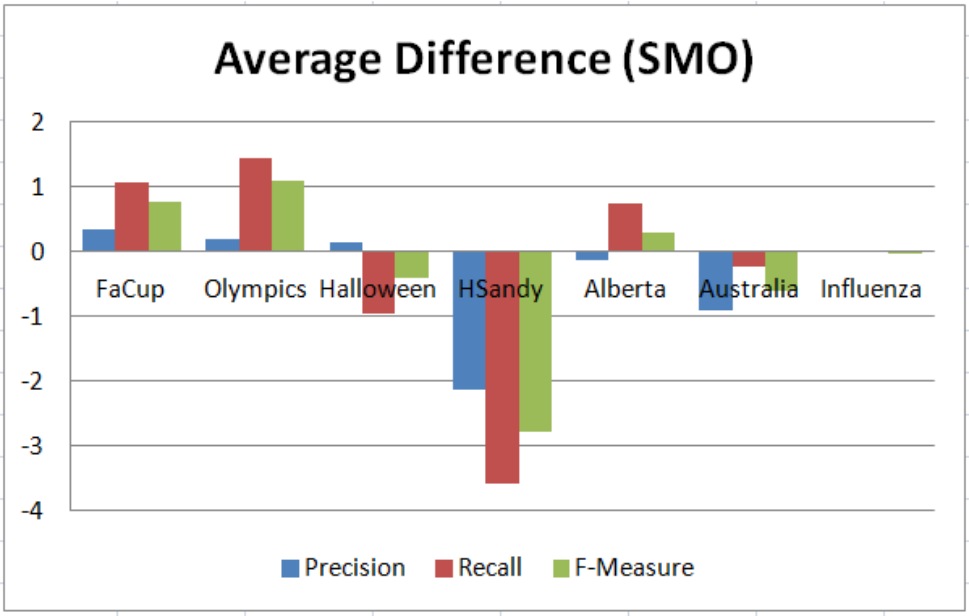
Source: the author.

Figure 5.10: Difference between the hybrid semantic and semantic-only enrichment configurations, in percentage points, using only the *CfsSubsetEval* algorithm, for the NB classifier



Source: the author.

Figure 5.11: Difference between the hybrid semantic and semantic-only enrichment configurations, in percentage points, using only the *CfsSubsetEval* algorithm, for the SMO classifier



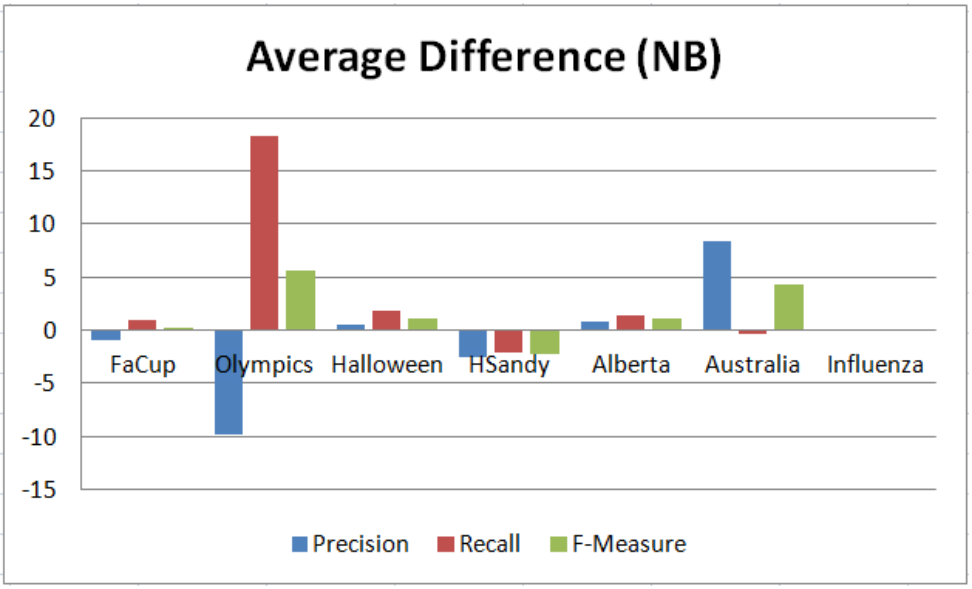
Source: the author.

5.4.3.2 Discussions

a) Qualitative and Quantitative Analysis of Conceptual Feature Extraction, Semantic Enrichment and Selection of Features

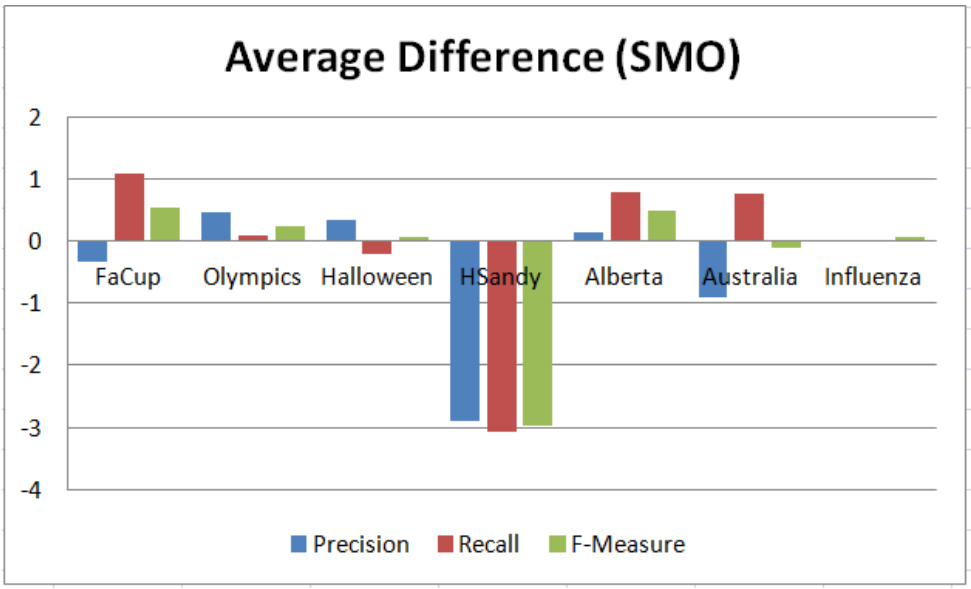
Considering the Conceptual Feature Extraction step, we observed that for extracting the conceptual features from URLs mentioned in the tweets (i.e. *hybrid semantic enrichment*), the

Figure 5.12: Difference between the hybrid semantic and semantic-only enrichment configurations in percentage points, in combination with the QUARTILES+CFS strategy, for the NB classifier



Source: the author.

Figure 5.13: Difference between the hybrid semantic and semantic-only enrichment configurations in percentage points, in combination with the QUARTILES+CFS strategy, for the SMO classifier



Source: the author.

Sports-related tweets presented more URLs associated with others tweets, whilst most tweets related to natural disasters and epidemics presented URLs to web documents with more relevant content (e.g. describing prevention measures, affected areas, and detailed reports). Some of these URLs no longer are valid, and thus were ignored by the analysis. This fact might have influenced our results, for in these cases, just fewer additional conceptual features were added

by the use of external documents (Table 5.11). In general, the adoption of external documents resulted in an increase of the named entities found, thus producing a bigger impact in the NER and ALL variations. Regarding the Semantic Enrichment step, DBpedia provided good coverage for finding resources related to the conceptual features through RapidMiner operators (80% in average).

We can also notice that the number features is drastically reduced for all target events and types of features when some technique for selecting discriminative features is applied (Table 5.11). However, the number of features were very similar regardless the type of enrichment, sometimes even equal. Thus we performed a manual analysis to observe whether the set of selected features were different.

Considering the application of *CfsSubsetEval* algorithm, we observed that for two datasets (FaCup and Olympics), the same set of features was selected for the TERMS variation (i.e. *semantic-only* and *hybrid semantic* enrichment). This is probably explained by the fact that no additional relevant term was added by the external documents. Otherwise, considering the combination of all types of features (i.e. ALL), both configurations presented very similar textual and semantic features. This is explained by the added number of named entities. In general, the *hybrid semantic enrichment* datasets presented more semantic features than the *semantic-only enrichment* ones, and in both cases the textual and semantic features in the top of the list are very related to the event analyzed.

With regard to the combined application of pruning and feature selection, we observed that except for the two datasets (FaCup and Olympics, for which the TERMS set of features were identical), the resulting set of features for each enrichment configuration was very different. However, we observed that the adoption of pruning increased the number of semantic features in the final feature set, particularly for the *hybrid semantic enrichment*.

b) Analysis of the Effect of the Different Methods for Selecting Features, Considering Semantic-only Enrichment

In Table 5.12, we present the results corresponding to the *semantic-only enrichment* configuration using *CfsSubsetEval* feature selection. We can observe that this approach was able to statistically outperform the baseline in 22.5% of the cases. In general, improvements could be noticed in 59.5% of the cases, mainly for the recall metric and the ALL variation. The HSandy was the dataset that statistically outperformed the baseline more times, with a maximum improvement of 10.6 pp for the TERMS type of feature and the Recall metric.

Table 5.13 presents the results for the *semantic-only enrichment* analysis according to

the combination of the pruning and *CfsSubsetEval* algorithms. We can observe that this approach statistically outperformed the baseline in 13.5% of the cases, most of them using the TERMS variation and considering the Recall metric. A greater number of statistically significant improvements could be noticed for the HSandy dataset, while for the Influenza and Halloween datasets, no statistical difference was noticed. In general, improvements could be noticed in 44.4% of cases, ranging from 0.1 pp to 7.7 pp.

In summary, better results were achieved by semantic-only enrichment when applying only the *CfsSubsetEval* algorithm, without performing the pruning step. Once no external conceptual features are incorporated to the list of features to be enriched, the number of resulting semantic features tend to be much smaller. Therefore, applying only the general-purpose feature selection algorithm is enough to achieve good results.

c) Performance Analysis of the Different Enrichment Configurations According to the Method for Selecting Features

The boxplots on Figures 5.8 and 5.9 enables the comparison of each enrichment configuration according to the applications of the feature selection techniques, considering the F-Measure metric.

For both techniques, the best results were achieved by the *hybrid semantic enrichment* configuration. In both cases, the presence of named entities helped to produce good results, reaching maximum values of 5.4 pp and 8.7 pp, for CFS and QUARTILES+CFS, respectively. Due to the lower volume of semantic features, the application of the QUARTILES+CFS combination produced inferior results than the CFS for the *semantic-only enrichment* configuration.

Using the *CfsSubsetEval* algorithm only, the *hybrid semantic enrichment* configuration achieved median improvements of 0.1 pp, 1.25 pp, and 0.7 pp for TERMS, NER, and ALL, respectively. Considering the maximum value, the variations achieved 2.8 pp, 5.4 pp, and 3.8 pp of improvement, for TERMS, NER, and ALL, respectively. For the *semantic-only enrichment* configuration, the ALL variation produced the best results with a median improvement of 0.5 pp, and 1.7 pp for the upper quartile.

For the combination of pruning and the *CfsSubsetEval* algorithm, the *hybrid semantic enrichment* configuration produced improvements of 0.25 pp, 1.62 pp, and 2.25 pp for the upper quartiles in TERMS, NER, and ALL, respectively. Regarding the maximum value, we achieved 3.1 pp for TERMS and 8.7 pp for NER and ALL variations. The *semantic-only enrichment* configuration was able to outperform the *hybrid semantic enrichment* only for the TERMS variation, which means that the elevate number of frequent and representative terms (Table

5.11) helped to improve the results.

d) 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. We detailed these results for each target event, event and technique for selecting discriminative features. In Figures 5.10 and 5.11, we analyze the performance considering only the application of *CfsSubsetEval* algorithm, and in Figures 5.12 and 5.13, we considered the combination of the pruning algorithm with the *CfsSubsetEval* algorithm.

Considering the application of the *CfsSubsetEval* algorithm only and the NB classification algorithm (Figure 5.10), the *hybrid semantic enrichment* configuration produced better results than the *semantic-only enrichment*, in terms of Precision, at the expense of Recall, and a small decrease could be observed in the F-Measure metric. For the Halloween and Influenza datasets, the difference between both strategies was minimum.

For the SMO classification algorithm (Figure 5.11), the *hybrid semantic enrichment* configuration produced similar patterns for the sportive events datasets (i.e. FaCup and Olympics), with slight improvement in Precision and a moderate improvement in terms of Recall. The natural disasters datasets HSandy, Australia Bushfire and Influenza achieved the worst results when the *hybrid semantic enrichment* configuration was applied in combination with the *CfsSubsetEval* algorithm.

In general, when only the *CfsSubsetEval* algorithm is applied, we could observe that the *hybrid semantic enrichment* configuration achieved better results in 32.5% of the cases, in comparison to the *semantic-only enrichment* configuration. In 23.8% of the cases no difference between the configurations could be noticed. For specific cases, the *hybrid semantic enrichment* configuration achieved improvements about 11.5 pp compared to the *semantic-only enrichment* configuration.

Considering the combination of pruning and the *CfsSubsetEval* algorithm (i.e. QUARTILES+CFS), for the NB algorithm, the sportive events FaCup and Olympics presented the same patterns, with improvements in Recall and F-Measure metrics, as presented in Figure 5.12. The difference for the Halloween and Alberta datasets was marginal, in both configurations. The HSandy dataset presented a small loss when submitted to the *hybrid semantic enrichment* configuration. In general, most of the improvements could be noticed in terms of Recall, different from the results achieved when using the *CfsSubsetEval* algorithm only. Therefore, pre-selecting the most discriminative semantic features helped to improve the Recall metric.

For the SMO algorithm, Figure 5.13, the difference was marginal for almost all datasets. HSandy dataset presented a decrease of 3 pp for F-Measure. No difference could be noticed for the Influenza dataset.

In general, we could observe that the *hybrid semantic enrichment* configuration achieved better results in 41.2% of the cases, in comparison to the *semantic-only enrichment* configuration. In 30.1% of the cases no difference between the configurations could be noticed. For specific cases, the *hybrid semantic enrichment* configuration achieved improvements about 29.3 pp compared to the *semantic-only enrichment* configuration.

In summary, *semantic-only* and *hybrid semantic* enrichment presented good results for event classification in tweets. For the *hybrid semantic enrichment* configuration, its combination with pruning and the *CfsSubsetEval* algorithm, produced the best results. Considering the *semantic-only enrichment*, in which the number of semantic features in smaller due to the lack of additional conceptual features provided by the external document enrichment, the best results were achieved using only the *CfsSubsetEval* algorithm. Regarding the classification algorithm, most of the expressive improvements could be noticed for the NB algorithm compared to the SMO algorithm. Among the features, the combination of NER and TERMS (i.e. ALL variation) provided overall the best results. Therefore, the *semantic-only enrichment* approach can be employed in situations where it is known that the dataset presents few URLs or the tweets are too old that the URLs are not available anymore. Otherwise, the *hybrid semantic enrichment* should be used, since it presented better results than the *semantic-only enrichment*, these improvements are more evident when considering the statistical test.

e) Final Comparative Analysis

In Table 5.14, we present a summarization of the experiments performed for the *semantic-only* and *hybrid semantic* enrichment configurations, considering the *CfsSubsetEval* feature selection algorithm and its combination with the pruning algorithm. According to the information presented in rows Without Pruning and CFS only, the *semantic-only* enrichment produced a larger number of improvements in comparison to the *hybrid semantic enrichment* configuration. However, considering the statistical analysis, the results produced by the latter outperformed the baseline in a greater number of times. The maximum value of improvement achieved was the same in both strategies (i.e. Precision metric and TERMS variation), considering the Without Pruning row. As resented in the *QUARTILES+CFS row*, the *hybrid semantic enrichment* configuration performed better in almost all situations, with improvements reaching almost five times more than the achieved in *semantic-only enrichment* configuration.

Table 5.14: Summarization of the results for semantic-only (SOE) and hybrid semantic enrichment (HSE), considering the *CfsSubsetEval* algorithm and its combination with pruning.

Configuration	Type of Enric.	General improvement	Statistically superior	Statistically inferior	Minimum improvement	Maximum improvement	Most improved type of feature	Most improved metric	Most improved dataset
Without Pruning	SOE	31.7%	07.1%	19.8 %	0.1 pp	6.9 pp	TERMS	Recall	Alberta F., Australia B., HSandy
	HSE	31.0%	12.7%	31.0%	0.1 pp	6.9 pp	NER and ALL	Recall and F-Measure	Alberta Floods
CFS only	SOE	59.5%	22.2%	15.0%	0.1 pp	10.6 pp	ALL	Recall	FaCup, HSandy, Australia B.
	HSE	52.3%	28.5%	11.0%	0.1 pp	8.7 pp	NER	Recall	FaCup, Olympics, Australia B., Alberta F.
QUARTILES+CFS	SOE	44.4%	13.5%	09.5%	0.1 pp	7.2 pp	TERMS	Recall	FaCup and HSandy
	HSE	53.17%	25.4%	12.7%	0.1 pp	32.6 pp	ALL	Recall	FaCup, Australia B., Alberta F.

Source: the author.

Considering the classification algorithms, most of the statistically superior and inferior results were achieved using the NB algorithm, for the *semantic-only enrichment* configuration. Regarding the event datasets, *semantic-only enrichment* produced excellent results for the HSandy dataset, as already reported in Figures 5.10, 5.11, 5.12, and 5.13.

In summary, to complement the information about the event by incorporating conceptual features from external document produced better results than using only the conceptual features from the tweet text. The pruning algorithm was able to handle this increase in the number of semantic features, as well as the application of the *CfsSubsetEval* algorithm only produced satisfactory results for the *semantic-only enrichment*.

5.5 Experiment #2: Hybrid Semantic Enrichment vs. Word Embeddings Approach

By executing this second experiment, we aim at evaluating the performance of the proposed framework (Figure 5.1) against an alternative form of enrichment, based on word embeddings. Thus, the process adopted includes hybrid enrichment, semantic feature pruning and general-purpose feature selection algorithm. The datasets employed in this analysis are the same used in *Experiment #1* (i.e. TERMS/NER/ALL, QUARTILES+CFS, NB and SMO classification algorithms).

5.5.1 Building the Baseline

Word embeddings is a distributional semantic approach, which produces word vectors for each word in the vocabulary. We employed a pre-trained word vectors using GloVe[8], produced over 2 billions tweets, representing a 1.2 million vocabulary. This vocabulary has terms from different languages.

To combine these word embeddings with the tweets of the target datasets, we employed the mean of the individual term's vector (LIU et al., 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 the word vectors. Then, we calculate the mean of all word vectors for this tweet. This aggregation allows a condensed embedding-based features representation, in which each tweet is represented by a unique vector, containing a 100-dimensional array.

These steps were applied for each target event dataset, using the Gensim[9] Python library. For the classification, we employed the supervised algorithms NB and SMO implementations available in this environment, adopting 10-fold cross-validation configuration.

5.5.2 Results

In Table 5.15, we present the results for the event classification task using the word embeddings approach in comparison to the application of the hybrid semantic enrichment framework, combined with the pruning algorithm and the *CfsSubsetEval* algorithm. These results represents one iteration of a 10-fold cross-validation configuration, considering the weighted F-Measure, Precision, and Recall metrics, respectively. As aforementioned, to train the models using the word embeddings we used the Python environment, from which we could not extract the result for the positive class only, as in the previous experiment.

In Figures 5.14 and 5.15, we present the average difference between our hybrid semantic enrichment framework for event classification in tweets and the approach using word embeddings, which have been widely used in NLP applications. The figures show the average difference for each target event dataset, considering the NB and SMO classification algorithms.

Finally, we validate our results for each metric through a statistical test, using two-tail paired $t - test$. For the comparison, we analyzed group of results (i.e. each dataset variation and classifier against the baseline built using word embeddings), using the Microsoft Excel. We

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

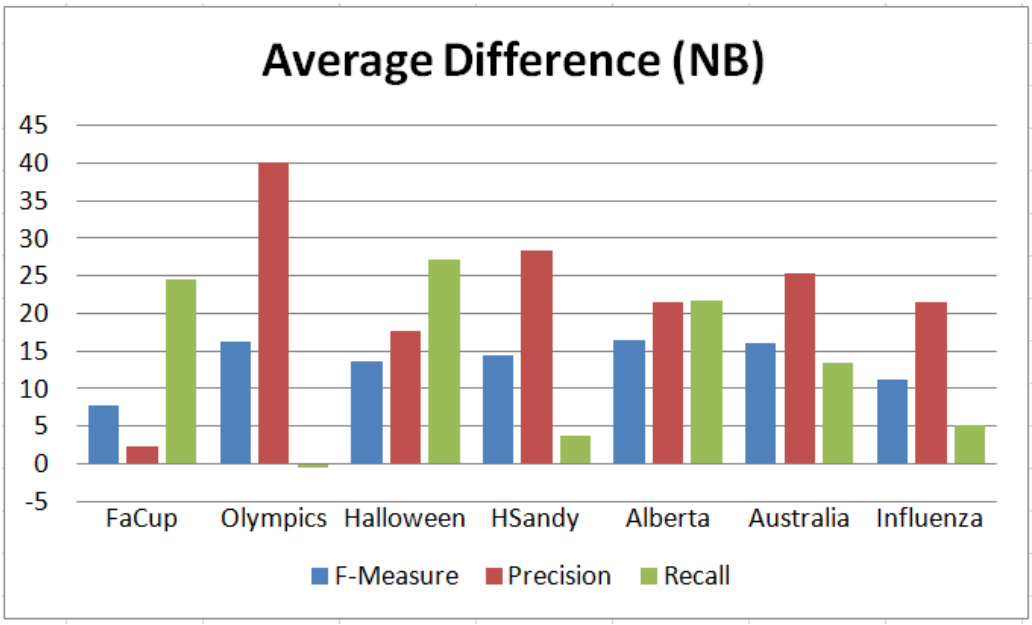
claim the improvement is *significant* with significance level of $\alpha = 0.05$, and *very significant* if $\alpha = 0.01$. Tables 5.16 and 5.17 summarize the result of our statistical analysis, for each classification algorithm.

Table 5.15: Comparison between the event classification in tweets using word embeddings against the hybrid semantic enrichment framework

Dataset	Algor.	word embeddings			TERMS			NER			ALL		
		F	P	R	F	P	R	F	P	R	F	P	R
FaCup	NB	0.831	0.891	0.666	0.915	0.920	0.918	0.902	0.910	0.905	0.912	0.915	0.914
	SMO	0.920	0.961	0.830	0.928	0.933	0.930	0.926	0.931	0.928	0.929	0.933	0.931
Olympics	NB	0.663	0.447	0.835	0.810	0.840	0.833	0.833	0.851	0.827	0.833	0.851	0.827
	SMO	0.820	0.949	0.575	0.883	0.893	0.891	0.884	0.899	0.893	0.884	0.899	0.893
Halloween	NB	0.743	0.704	0.607	0.874	0.875	0.874	0.882	0.883	0.881	0.882	0.883	0.881
	SMO	0.799	0.867	0.612	0.901	0.901	0.900	0.905	0.905	0.904	0.905	0.905	0.904
HSandy	NB	0.774	0.634	0.881	0.923	0.923	0.923	0.912	0.913	0.913	0.919	0.920	0.920
	SMO	0.864	0.906	0.735	0.941	0.941	0.942	0.931	0.934	0.933	0.934	0.936	0.935
Alberta Floods	NB	0.831	0.781	0.778	0.988	0.988	0.988	0.999	0.999	0.999	0.999	0.999	0.999
	SMO	0.862	0.891	0.747	0.996	0.996	0.996	0.999	0.999	0.999	0.999	0.999	0.999
Australia Bushfire	NB	0.824	0.733	0.851	0.978	0.979	0.978	0.989	0.989	0.989	0.989	0.989	0.989
	SMO	0.827	0.833	0.706	0.990	0.990	0.990	0.993	0.993	0.993	0.995	0.995	0.995
Influenza	NB	0.887	0.784	0.948	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	SMO	0.978	0.983	0.959	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

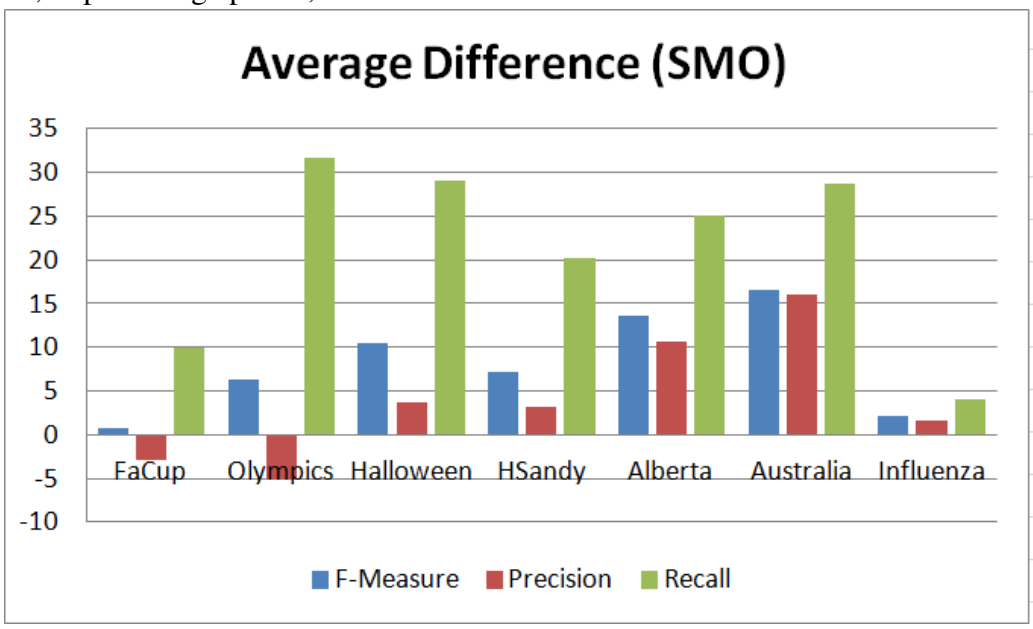
Source: the author.

Figure 5.14: Difference between the Hybrid Semantic Enrichment and the Word Embeddings approach, in percentage points, for the NB classifier



Source: the author.

Figure 5.15: Difference between the Hybrid Semantic Enrichment and the Word Embeddings approach, in percentage points, for the SMO classifier



Source: the author.

Table 5.16: Statistical $t - test$ for the NB classifier

Type of Feature	F-Measure	Precision	Recall
TERMS	0.00428	0.00240	0.02190
NER	0.00307	0.00221	0.02373
ALL	0.00261	0.00206	0.02161

Source: the author.

Table 5.17: Statistical $t - test$ for the SMO classifier

Type of Feature	F-Measure	Precision	Recall
TERMS	0.01832	0.18299	0.00152
NER	0.01898	0.17073	0.00153
ALL	0.01763	0.16229	0.00147

Source: the author.

5.5.3 Discussions

a) Word Embeddings for Event Classification in Tweets

In Table 5.15, we observe that the word embeddings approach produces significant results for the event classification task, but not for all target events tested, considering the weighted F-Measure and the NB classification algorithm. For the SMO algorithm, the results were slightly better. Notice that the former algorithm produces better results in terms of Recall, whereas Precision is the best metric for the latter. This difference in the performance affects the results discussed in this section.

Considering the word embeddings model as the baseline for our hybrid semantic enrichment configuration datasets, we could observe improvements in 95.2% of the cases, ranging from 0.6 pp to 40.4 pp in specific situations (i.e. the Olympics dataset, for NER and ALL variations, considering the Precision metric). For both classification algorithms, the average improvement was about 12.04 pp, 16.07 pp, and 15.56 pp, for F-Measure, Precision, and the Recall metric.

b) Classification Performance Comparison

In Figures 5.14 and 5.15, we compared the difference between our approach and the word embeddings baseline. Considering the NB classifier (Figure 5.14), for the natural disaster datasets (i.e. HSandy, Alberta Floods, and Australia Bushfire datasets), the difference was higher in terms of Precision, ranging from 21.4 pp to 28.4 pp. Our approach achieved improvements of 40.4 pp for the Olympics dataset, considering the Precision metric. Great results were achieved in Recall, for Halloween and FaCup datasets, with 27.1 pp and 24.6 pp of difference, respectively. The Influenza dataset present good results, mainly for Precision with an average improvement of 21.5 pp. This difference between the Precision and Recall metrics is partially explained by the baseline performance, since the Precision metric was not always good.

For the SMO classifier (Figure 5.15), all datasets performed better than the word embeddings approach in terms of Recall, ranging from 4 pp to 31 pp of improvement. Using embeddings, the classification resulted in an improvement of Precision for the sportive events FaCup and Olympics. The natural disaster datasets presented similar patterns, with more pronounced improvement in terms of Recall, followed by F-Measure. The Influenza dataset resulted in a small improvement in all metrics. Likewise, the difference between the Precision and Recall metrics is partially explained by the performance of Recall metric on the respective baseline.

Tables 5.16 and 5.17 summarize the result of our statistical analysis. For the NB classifier, all types of features achieved very significant improvements for F-Measure and Precision metrics (i.e. $alpha = 0.01$). For the Recall metric, the improvements were significant (i.e. $alpha = 0.05$). As shown in Table 5.17, for the SMO classifier the improvements were very significant in all types of features for Recall, and significant for F-Measure metric.

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 two distinct algorithms widely used for text classification, and outperformed the results achieved using a word embeddings approach, which has been used in application related to text classification, topical clustering, and question answering.

6 CONCLUSION AND FUTURE WORKS

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 source enrichment and named entity extraction through NER tools. 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 the most discriminative features resulting from this process, using two complementary techniques: a specific-purpose pruning algorithm and general-purpose feature selection algorithms. These elements were evaluated in a broad experimental setting. The proposed approach does not rely on assumptions about the types of events, 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.

Regarding the contextual enrichment techniques proposed: a) applying the NER tools helped to increase the classification performance since most of the improved results were achieved in the variations composed of named entities (i.e. NER and ALL); b) the external document enrichment contributed with more information about the event at hand, particularly new vocabulary (i.e. frequent and representative terms) not detected in the target event dataset, due to the sparse nature of tweets; and c) the semantic enrichment using a LOD cloud knowledge base helped to generalize the information about the event with useful knowledge, although the selection of the relevant semantic features resulting from the enrichment is an issue that can degrade the classification performance if not carefully handled.

For the Semantic Enrichment step, the DBpedia knowledge base presented a good coverage (i.e. 80% in average) for enriching the conceptual features extracted. Regarding the property analyzed, the *rdf:type* provided useful information about the event improving event classification, but only this specific property was explored in the current work. According to the domain of the event, other LOD cloud datasets could also be employed, as well as other properties, a topic that deserves further investigation.

The specific-purpose pruning algorithm did improve the presence of relevant semantic features in the training dataset, but when applied alone the improvements on the classification results were marginal. Its combination with feature selection algorithms, specifically the *CfsSubsetEval*, produced much better results, but it depends on the characteristics of the classification algorithm applied. For instance, the SMO classification algorithm is less sensitive to the huge volume of features compared to NB, and thus the pruning effects were less significant.

In the evaluation experiments, just two classification algorithms were employed. Considering the related work, there are other classification algorithms that presented good results for text classification in tweets (e.g. Random Forest, JRip, and Maximum Entropy), which can be explored in future work.

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. Considering the textual features baseline, it achieved improvements in 53.17% of the cases, whereas the improvements could be noticed in 95.2% of the cases for the baseline using word embeddings. Despite the promising results in datasets representing events of 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.

The results achieved during this research so far resulted into two publications:

- ROMERO, S. A. P.; BECKER, K. Experiments with semantic enrichment for event classification in tweets. In: **Proc. of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence**. Omaha, Nebraska, USA: [s.n.], 2016.

- ROMERO, S. A. P.; BECKER, K. A semantic enrichment framework for classifying events in tweets. In: **Proc. of the 2016 SBBD/WTDBD - XV Workshop of Theses and Dissertations in Databases**. Salvador, Bahia, BRA: [s.n.], 2016.

In Table 6.1, we summarize the main contribution of each publication:

Table 6.1: Comparison to related work

Work	Description	Features			Selection of features	Learning Technique	Event Type
		Tweet	External	Semantic			
(ROMERO; BECKER, 2016a)	Semantic enrichment framework for event classification in tweets	Named entities, frequent and representative terms	NO	DBpedia (rdf:type)	general-purpose	NB and SMO	General (specified, planned and unplanned)
(ROMERO; BECKER, 2016b)	Hybrid semantic enrichment framework for event classification in tweets	Named entities, frequent and representative terms	Analysis of the URL content	DBpedia (rdf:type)	general-purpose and semantic feature pruning	NB and SMO	General (specified, planned and unplanned)

Source: the author.

Given these results and limitations, future work involves: a) experimenting with other properties and knowledge bases available in the LOD cloud; b) improving the techniques to select more discriminative features; c) define an architecture based on hybrid semantic enrichment that encompasses the Event Identification and Classification in tweets, according to the process defined; d) elaborate an experiment that allows us to compare our approach with other state-of-the-art event classification methods; e) elaborate an approach to cluster similar events that occur in different places and periods; f) adapt the framework to identify the events on a more general level, for example, instead of identifying tweets related to a specific event type (i.e. the final football season in England), focus on a general domain (i.e. sportive events); g) identify other event target datasets to be used in the experiments; h) apply other classification algorithms.

REFERENCES

ABEL, F. et al. Semantics + Filtering + Search = Twitcident. Exploring Information in Social Web Streams. In: **Proceedings of the 23rd ACM Conference on Hypertext and Social Media**. New York, NY, USA: ACM, 2012. (HT '12), p. 285–294. <http://doi.acm.org/10.1145/2309996.2310043>. Accessed in: 2017-03-21.

ABEL, F. et al. Twitcident: Fighting Fire with Information from Social Web Streams. In: **Proceedings of the 21st International Conference on World Wide Web**. New York, NY, USA: ACM, 2012. (WWW '12 Companion), p. 305–308. <http://doi.acm.org/10.1145/2187980.2188035>. Accessed in: 2017-03-21.

ABELLÓ, A. et al. Using Semantic Web Technologies for Exploratory OLAP: A Survey. **IEEE Transactions on Knowledge and Data Engineering**, v. 27, n. 2, p. 571–588, 2015.

AIELLO, L. et al. Sensing trending topics in Twitter. **IEEE Transactions on Multimedia**, IEEE, v. 15, n. 6, p. 1–15, 2013.

ANANTHARAM, P. et al. Extracting City Traffic Events from Social Streams. **ACM Transactions on Intelligent Systems and Technology**, ACM, New York, NY, USA, v. 6, n. 4, p. 43:1–43:27, 2015. <http://doi.acm.org/10.1145/2717317>. Accessed in: 2017-03-21.

ARAMAKI, E.; MASKAWA, S.; MORITA, M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In: **Proceedings of the Conference on Empirical Methods in Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (EMNLP '11), p. 1568–1576. <http://dl.acm.org/citation.cfm?id=2145432.2145600>. Accessed in: 2017-03-21.

ATEFEH, F.; KHREICH, W. A Survey of Techniques for Event Detection in Twitter. **Computational Intelligence**, Blackwell Publishers, Inc., Cambridge, MA, USA, v. 31, n. 1, p. 132–164, 2015. <http://dx.doi.org/10.1111/coin.12017>. Accessed in: 2017-03-21.

AUER, S. et al. DBpedia: A Nucleus for a Web of Open Data. In: **Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference**. Berlin, Heidelberg: Springer-Verlag, 2007. (ISWC'07/ASWC'07), p. 722–735. <http://dl.acm.org/citation.cfm?id=1785162.1785216>. Accessed in: 2017-03-21.

BECKER, H. et al. Identifying Content for Planned Events Across Social Media Sites. In: **Proceedings of the Fifth ACM International Conference on Web Search and Data Mining**. New York, NY, USA: ACM, 2012. (WSDM '12), p. 533–542. <http://doi.acm.org/10.1145/2124295.2124360>. Accessed in: 2017-03-21.

BECKER, H.; NAAMAN, M.; GRAVANO, L. Learning Similarity Metrics for Event Identification in Social Media. In: **Proceedings of the Third ACM International Conference on Web Search and Data Mining**. New York, NY, USA: ACM, 2010. (WSDM '10), p. 291–300. <http://doi.acm.org/10.1145/1718487.1718524>. Accessed in: 2017-03-21.

BECKER, H.; NAAMAN, M.; GRAVANO, L. Beyond trending topics: Real-world event identification on Twitter. In: **International AAAI Conference on Web and Social Media**. Barcelona, Catalonia, Spain: [s.n.], 2011.

BERNERS-LEE, T.; HENDLER, J.; ORA, L. The Semantic Web. **Scientific American**, p. 29–37, 2001.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data - the story so far. **International Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1–22, 2009.

BIZER, C. et al. DBpedia - A Crystallization Point for the Web of Data. **Web Semant.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 7, n. 3, p. 154–165, 2009. <http://dx.doi.org/10.1016/j.websem.2009.07.002>. Accessed in: 2017-03-21.

CALLEGARI-JACQUES, S. **Bioestatística: Princípios e aplicações**. [S.l.]: Artmed Editora, 2009. <https://books.google.com.br/books?id=nuaVLSCiAgsC>. Accessed in: 2017-03-21.

CAMBRIA, E. et al. New Avenues in Opinion Mining and Sentiment Analysis. **IEEE Intelligent Systems**, IEEE Computer Society, Los Alamitos, CA, USA, v. 28, n. 2, p. 15–21, 2013.

FISICHELLA, M. et al. Detecting health events on the social web to enable epidemic intelligence. In: **Proceedings of the 18th International Conference on String Processing and Information Retrieval**. Berlin, Heidelberg: Springer-Verlag, 2011. (SPIRE'11), p. 87–103. <http://dl.acm.org/citation.cfm?id=2051073.2051083>. Accessed in: 2017-03-21.

GENC, Y.; SAKAMOTO, Y.; NICKERSON, J. V. Discovering context: Classifying tweets through a semantic transform based on wikipedia. In: **Proceedings of the 6th International Conference on Foundations of Augmented Cognition: Directing the Future of Adaptive Systems**. Berlin, Heidelberg: Springer-Verlag, 2011. (FAC'11), p. 484–492. <http://dl.acm.org/citation.cfm?id=2021773.2021833>. Accessed in: 2017-03-21.

HALL, M. et al. The weka data mining software: an update. **ACM SIGKDD Explorations Newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

JANPUANGTONG, S.; SHELL, D. A. Leveraging Ontologies to Improve Model Generalization Automatically with Online Data Sources. In: BONET, B.; KOENIG, S. (Ed.). **Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.** [S.l.]: AAAI Press, 2015. p. 3981–3986. <http://www.aaai.org/ocs/index.php/IAAI/IAAI15/paper/view/9486>. Accessed in: 2017-03-21.

KENTER, T.; RIJKE, M. de. Short Text Similarity with Word Embeddings. In: **Proceedings of the 24th ACM International on Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2015. (CIKM '15), p. 1411–1420. <http://doi.acm.org/10.1145/2806416.2806475>. Accessed in: 2017-03-21.

KHUC, V. N. et al. Towards Building Large-scale Distributed Systems for Twitter Sentiment Analysis. In: **Proceedings of the 27th Annual ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2012. (SAC '12), p. 459–464. <http://doi.acm.org/10.1145/2245276.2245364>. Accessed in: 2017-03-21.

LAMB, A.; PAUL, M. J.; DREDZE, M. Separating fact from fear: Tracking flu infections on twitter. In: **North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)**. [S.l.: s.n.], 2013. <http://www.aclweb.org/anthology/N/N13/N13-1097.pdf>. Accessed in: 2017-03-21.

LI, C.; SUN, A.; DATTA, A. Twevent: Segment-based Event Detection from Tweets. In: **Proceedings of the 21st ACM International Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2012. (CIKM '12), p. 155–164. <http://doi.acm.org/10.1145/2396761.2396785>. Accessed in: 2017-03-21.

LI, Q. et al. Tweet topic classification using distributed language representations. In: **Proc. of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence**. Omaha, Nebraska, USA: [s.n.], 2016.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 17, n. 4, p. 491–502, abr. 2005. <http://dx.doi.org/10.1109/TKDE.2005.66>. Accessed in: 2017-03-21.

LIU, X. et al. Reuters Tracer: A Large Scale System of Detecting & Verifying Real-Time News Events from Twitter. In: **Proceedings of the 25th ACM International on Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2016. (CIKM '16), p. 207–216. <http://doi.acm.org/10.1145/2983323.2983363>. Accessed in: 2017-03-21.

LIU, Y. et al. Topical Word Embeddings. In: **Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence**. [S.l.]: AAAI Press, 2015. (AAAI'15), p. 2418–2424. <http://dl.acm.org/citation.cfm?id=2886521.2886657>. Accessed in: 2017-03-21.

MCMINN, A. J.; MOSHFEGHI, Y.; JOSE, J. M. Building a Large-scale Corpus for Evaluating Event Detection on Twitter. In: **Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management**. New York, NY, USA: ACM, 2013. (CIKM '13), p. 409–418. <http://doi.acm.org/10.1145/2505515.2505695>. Accessed in: 2017-03-21.

MEDVET, E.; BARTOLI, A. Brand-Related Events Detection, Classification and Summarization on Twitter. In: **Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01**. Washington, DC, USA: IEEE Computer Society, 2012. (WI-IAT '12), p. 297–302. <http://dl.acm.org/citation.cfm?id=2457524.2457698>. Accessed in: 2017-03-21.

MIKOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space. **CoRR**, abs/1301.3, 2013. <http://arxiv.org/abs/1301.3781>. Accessed in: 2017-03-21.

MOHAMMAD, S. M.; KIRITCHENKO, S.; ZHU, X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. **CoRR**, abs/1308.6, 2013. <http://arxiv.org/abs/1308.6242>. Accessed in: 2017-03-21.

NILSSON, R. et al. Evaluating Feature Selection for SVMs in High Dimensions. In: **Machine Learning: ECML 2006 SE - 72**. [S.l.]: Springer Berlin Heidelberg, 2006, (Lecture Notes in Comp. Science, v. 4212). p. 719–726.

NOURBAKHSH, A. et al. Newsworthy Rumor Events: A Case Study of Twitter. **2015 IEEE International Conference on Data Mining Workshop (ICDMW)**, IEEE Computer Society, Los Alamitos, CA, USA, v. 00, n. undefined, p. 27–32, 2015.

OLTEANU, A.; VIEWEG, S.; CASTILLO, C. What to expect when the unexpected happens: Social media communications across crises. In: **Proc. of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing**. New York, NY, USA: ACM, 2015. (CSCW '15), p. 994–1009. <http://doi.acm.org/10.1145/2675133.2675242>. Accessed in: 2017-03-21.

PACKER, H. S. et al. Event Detection Using Twitter and Structured Semantic Query Expansion. In: **Proceedings of the 1st International Workshop on Multimodal Crowd Sensing**. New York, NY, USA: ACM, 2012. (CrowdSens '12), p. 7–14. <http://doi.acm.org/10.1145/2390034.2390039>. Accessed in: 2017-03-21.

PAGE, L. et al. **The PageRank Citation Ranking: Bringing Order to the Web.** [S.l.], 1999. <http://ilpubs.stanford.edu:8090/422/>. Accessed in: 2017-03-21.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. GloVe: Global Vectors for Word Representation. In: **Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. <http://www.aclweb.org/anthology/D14-1162>. Accessed in: 2017-03-21.

PETROVIC, S. et al. Can Twitter Replace Newswire for Breaking News? In: KICIMAN, E. et al. (Ed.). **ICWSM**. [S.l.]: The AAAI Press, 2013.

PLATT, J. **Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines**. [S.l.], 1998. 21 p. <http://bit.ly/2nsjf9K>. Accessed in: 2017-03-21.

REUTER, T.; CIMIANO, P. Event-based classification of social media streams. In: **Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval**. New York, NY, USA: ACM, 2012. (ICMR '12), p. 22:1–22:8. <http://doi.acm.org/10.1145/2324796.2324824>. Accessed in: 2017-03-21.

ROMERO, S. A. P.; BECKER, K. Experiments with semantic enrichment for event classification in tweets. In: **Proc. of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence**. Omaha, Nebraska, USA: [s.n.], 2016.

ROMERO, S. A. P.; BECKER, K. A semantic enrichment framework for classifying events in tweets. In: **Proc. of the 2016 SBBD/WTDBD - XV Workshop of Theses and Dissertations in Databases**. Salvador, Bahia, BRA: [s.n.], 2016.

ROSA, K. D. et al. Topical clustering of tweets. **Proc. of the ACM SIGIR: SWSM**, 2011.

ROWE, M.; STANKOVIC, M. Aligning Tweets with Events: Automation via Semantics. **Semantic Web Journal**, 2011. <http://www.semantic-web-journal.net/content/aligning-tweets-events-automation-semantics>. Accessed in: 2017-03-21.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995.

SAIF, H.; HE, Y.; ALANI, H. Semantic Sentiment Analysis of Twitter. In: CUDRÉ-MAUROUX, P. et al. (Ed.). **CEUR Workshop Proceedings**. [S.l.]: Springer Berlin Heidelberg, 2012. (Lecture Notes in Computer Science, v. 917), p. 56–66. <http://dx.doi.org/10.1007/978-3-642-35176-1_32>. Accessed in: 2017-03-21.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: **Proceedings of the 19th International Conference on World Wide Web**. New York, NY, USA: ACM, 2010. (WWW '10), p. 851–860. <http://doi.acm.org/10.1145/1772690.1772777>. Accessed in: 2017-03-21.

SANKARANARAYANAN, J. et al. TwitterStand: News in Tweets. In: **Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**. New York, NY, USA: ACM, 2009. (GIS '09), p. 42–51. <http://doi.acm.org/10.1145/1653771.1653781>. Accessed in: 2017-03-21.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the Linked Data Best Practices in Different Topical Domains. In: MIKA, P. et al. (Ed.). **Proceedings of the 13th International Semantic Web Conference**. [S.l.]: Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8796). p. 245–260. <http://dx.doi.org/10.1007/978-3-319-11964-9_16>. Accessed in: 2017-03-21.

SCHULZ, A.; GUCKELSBERGER, C.; JANSSEN, F. Semantic Abstraction for Generalization of Tweet Classification: An Evaluation on Incident-Related Tweets. **Semantic Web Journal**, p. 1–21, 2015.

SCHULZ, A.; RISTOSKI, P. The Car That Hit The Burning House: Understanding Small Scale Incident Related Information in Microblogs. In: **AAAI Technical Report / WS**. [S.l.]: AAAI Press, 2013. v. 13-04, n. 71.

SCHULZ, A.; RISTOSKI, P.; PAULHEIM, H. I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs. In: **The Semantic Web: ESWC 2013 Satellite Events**. Montpellier, France: Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 7955). p. 22–33.

SCHULZ, A.; SCHMIDT, B.; STRUFE, T. Small-Scale Incident Detection Based on Microposts. In: **Proceedings of the 26th ACM Conference on Hypertext & Social Media**. New York, NY, USA: ACM, 2015. (HT '15), p. 3–12. <http://doi.acm.org/10.1145/2700171.2791038>. Accessed in: 2017-03-21.

TECHENTIN, R. W. et al. Implementing Iterative Algorithms with SPARQL. In: **EDBT/ICDT Workshops**. [S.l.: s.n.], 2014. p. 216–223.

TSOU, M.-H. et al. Social Media Analytics and Research Test-bed (SMART Dashboard). In: **Proceedings of the 2015 International Conference on Social Media & Society**. New York, NY, USA: ACM, 2015. (SMSociety '15), p. 2:1–2:7. <http://doi.acm.org/10.1145/2789187.2789196>. Accessed in: 2017-03-21.

VOSECKY, J. et al. Dynamic Multi-faceted Topic Discovery in Twitter. In: **Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management**. New York, NY, USA: ACM, 2013. (CIKM '13), p. 879–884. <http://doi.acm.org/10.1145/2505515.2505593>. Accessed in: 2017-03-21.

VOSECKY, J. et al. Integrating Social and Auxiliary Semantics for Multifaceted Topic Modeling in Twitter. **ACM Transactions on Internet Technology**, ACM, New York, NY, USA, v. 14, n. 4, p. 27:1–27:24, 2014. <http://doi.acm.org/10.1145/2651403>. Accessed in: 2017-03-21.

Appendices

AppendixA

In this Appendix, we present the first experiments performed to analyze the contribution of the each type of core feature proposed, and whether they are related to specific types of events.

A.1 Motivation

In Chapter 4, we presented a set of core features selected to characterize an event, namely frequent and representative Vocabulary, Agents, and Location. Each one of them represents an important element of the event definition adopted by this work. As an attempt to understand the contribution of each type of core feature for event classification in tweets, we performed a set of experiments, in which we adopted four datasets representing events of distinct natures and different combinations of the proposed core features. By performing this experiment, we aim at analyzing:

- which combination of the core features produce better results for the event classification task;
- what is the impact of each feature in the classification task;
- what is the impact of each feature with regard to the different event types.

A.2 Experiment Description

The datasets used in our experiments were the same described in Chapter 5, namely FaCup, HSandy, Halloween, and Olympics[1]. We applied only semantic enrichment (i.e. without the adoption of external document enrichment), as presented in Figure A.1.

To compare the contribution of semantic enrichment of tweets according to the proposed framework against the traditional term-based classification, we prepared different mining datasets for each one of the target events:

- baseline: composed by terms extracted from tweets. The terms were extracted from the

[1]To produce these datasets, positive examples were extracted from the target event datasets. The negative examples in the dataset are composed by tweets extracted from the other three datasets in combination with tweets from the SemEval-Task4 dataset.

tweets using the filter *StringToWordVector* parametrized to generate all alphabetic uni-grams;

- fully enriched dataset: composed by the incorporation of the uni-grams extracted from the tweets and the semantic features resulting from the enrichment of Agents (A), Location (L), frequent terms (F) and domain representative terms (T). We refer to this combination as A_L_F_T;

- partially enriched datasets: these datasets were created to analyze the contribution of each one of the proposed types of core features in the classification of events, in a one-leave-out strategy. We incorporated the uni-grams extracted from tweets with combinations of three types of semantic features. The tested combinations were A_L_T (without frequent terms), A_L_F (without representative terms), A_T_F (without location), and L_T_F (without agents).

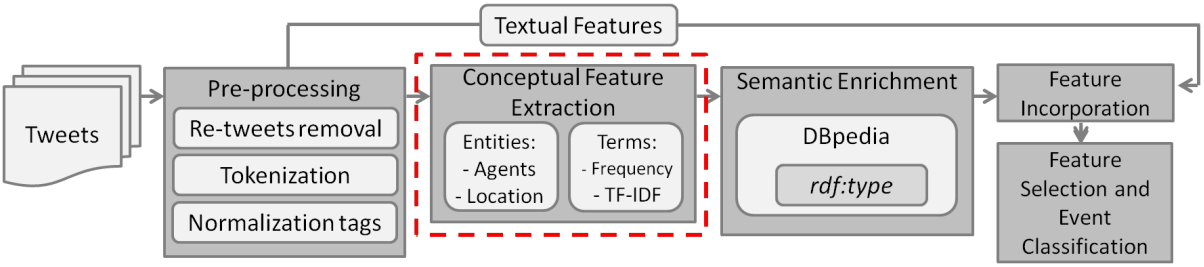
Considering the number of textual and semantic features resulting from the Incorporation step, we prepared two versions of each one of the aforementioned mining datasets: without feature selection, and with feature selection. As feature selection algorithm, we tested *CfsSubsetEval* and *InformationGain*, both available in Weka. The results reported in this Appendix refer to *CfsSubsetEval* algorithm, which were superior to the ones using the *InformationGain* algorithm.

We used two classification algorithms available in Weka, NB and SMO, using the default parameters, and a 10-fold cross-validation configuration. To statistically compare the results using the paired $t-test$, we used the Weka Experimenter parametrized with: a) a set of datasets reflecting the baseline and combinations of data enrichment, with and without feature selection; b) five (5) executions for each algorithm, using Weka's default parameters for these algorithms; and c) 10-fold cross-validation. Finally, we compared the results using the F-Measure, Precision and Recall metrics.

A.3 Dataset Preparation

We developed the experiments according to the steps presented in Figure A.1, an initial version of our proposed framework, in which external documents enrichment and semantic feature pruning were not explored. As highlighted in Figure A.1, the focus of this analysis is the Conceptual Feature Extraction step. The preparation of the datasets followed the same steps presented in Chapter 5.

Figure A.1: Summarized Pipeline for the Event Classification process



Source: the author

A.4 Results and Discussion

In this section, we describe the results of our experiments, the evaluation, and the statistical analysis.

A.4.1 Conceptual Feature Extraction

After the Pre-processing step, we extracted the core features Agents, Location, Frequent Terms and Representative Terms from each target dataset. Table A.1 presents the number of conceptual features extracted from each dataset using the Open Calais API. We could notice that for planned sportive events there are a smaller number of locations, compared to the other target events. The HSandy dataset presented the smallest number of agents.

Table A.1: Summary of the number of conceptual features extracted, representing agents and locations

Datasets	Conceptual Features	
	Agents	Location
FaCup	22	6
Olympics	71	17
Halloween	68	65
HSandy	11	57

Source: the author

Given the conceptual features extracted from the tweets, we created combinations of the features to observe the influence of each one of them in the Event Classification task. Thus, we

generated a list of all core features, namely A_L_T_F, and then we eliminate one by one resulting in other four core feature combinations: A_L_T, A_L_F, A_T_F, and L_T_F. Following this configuration, we created a set of twelve (12) datasets for each target event:

- the baseline;
- a fully enriched dataset (i.e. A_L_T_F);
- four partially enriched datasets (i.e. A_L_T, A_L_F, A_T_F, and L_T_F).

Each one of these datasets was created with and without feature selection.

A.4.2 Semantic Enrichment

The core feature combinations were used as input to the Semantic Enrichment step. For each target event, Table A.2 presents the number of semantic features resulting from the enrichment of the A_L_T_F combination.

The DBpedia knowledge base showed good coverage, enriching about 82% of the conceptual features submitted. Regarding the direct type property, it resulted in a significant number of features to be incorporated, but yielding quite sparse datasets.

Table A.2: Summary of the number of conceptual features submitted to DBPedia, number of matches, and number of Direct Types retrieved

#	Datasets			
	FaCup	Olympics	Halloween	HSandy
Submitted	50	96	152	83
Matched	40	68	140	70
Direct types	452	583	1020	454

Source: the author

A.4.3 Feature Selection

After the Semantic Enrichment step, we performed the incorporation of the semantic features resulting from the previous step and the textual features extracted from the tweets. Then, we prepared two configurations of each incorporated dataset, with and without feature selection, through the application of the *CfsSubsetEval* algorithm.

In Table A.3, we present the amount of textual and semantic features resulting from the Semantic Enrichment step without applying the feature selection algorithm (WP row) and the amount of textual and semantic features resulting from the application of the *CfsSubsetEval* algorithm (CFS row). In average, the application of the feature selection algorithm reduced the number of textual and semantic features in 96.42%.

Table A.3: Number of features resulting from the Semantic Enrichment step (WP) and the Feature Selection step (CFS), for the A_L_T_F combination

Configuration	Technique	Datasets			
		FaCup	Olympics	Halloween	HSandy
Baseline	WP	1672	1827	1892	2105
	CFS	64	103	151	75
A_L_T_F	WP	2104	2405	2844	2555
	CFS	71	90	131	65

Source: the author.

A manual analysis for the A_L_T_F combination for the FaCup dataset revealed that most of the semantic features presented in the top of the list are related to the event topic, while the textual features are extremely related to the domain analyzed. In the Olympics dataset most of the semantic features are related to locations whereas the textual features are very related to the domain of the event analyzed, mentioning teams, their members, and other terms related to the sports practiced. For the Halloween dataset, the application of the *CfsSubsetEval* algorithm resulted in a mix of textual and semantic features related to the domain of the event and to locations. Considering that in these experiments no semantic feature pruning was applied, the *owl#Thing* concept appeared in the top of the resulting list of the HSandy dataset, most of the other features are related to locations.

A.4.4 Event Classification

In the Classification step, we applied the NB and the SMO classification algorithms to all datasets. Table A.4 summarizes the results of our evaluation in terms of Precision (P), Recall (R), and F-Measure (F). We report the results of the positive class only, focus of our analysis, since we aim at identifying the tweets related to a specific event. For the NB classifier we considered the results using the filter for attribute selection *CfsSubsetEval*, whereas for the SMO, we decided to report just the original results (i.e. without *CfsSubsetEval* algorithm), once the SMO kernel performed better for this amount of textual and semantic features (NILSSON et al., 2006).

The results presented in Table A.4 show that semantic enrichment does improve the Event Classification in all datasets, considering the NB classifier. Significant improvement can be noticed on Recall for FaCup, Olympics and HSandy datasets, whereas for the Halloween dataset, the improvements were more expressive in terms of Precision. Considering the SMO classifier, the improvements were less significant and restricted to the FaCup, Olympics, and HSandy datasets.

Considering the feature combinations, A_L_T_F produced the best results for both classification algorithms and all target events. Secondly, the A_L_F and A_L_T combination produced results that outperformed the baseline in almost all target events, mainly for Recall and F-Measure metrics, which means that frequent (F) and representative terms (F) presented similar contribution for the classification. The absence of the location core feature (i.e. the A_F_T combination) produced good results for the FaCup and HSandy datasets, considering both algorithms.

In order to perform a deeper analysis, we applied a statistical test to verify the superiority of these results compared to the baseline. Tables A.5, A.6, and A.7 summarize the results of the two-tail paired $t - test$, with 0.05 significance, by comparing the Recall, F-Measure, and Precision metrics for each feature combination against the baseline, respectively. To produce this analysis, we combined the results achieved for the NB and SMO algorithms, considering the positive class. In these tables, the results depicted with a (*) represent that the baseline is statistically superior, the (v) symbol means that the combination analyzed is statically superior against the baseline. Otherwise, there is no statistic difference on the results.

As presented in Table A.5 (i.e. Recall metric), the use of semantically enriched features improved the ability to retrieve relevant tweets for the FaCup and HSandy datasets, presenting

Table A.4: Results for NB and SMO classification algorithms

Dataset	Feature Combination	NB (with feature selection)			SMO (without feature selection)		
		P	R	F	P	R	F
FaCup	Baseline	0.966	0.730	0.832	0.941	0.907	0.924
	A_L_F_T	0.915	0.775	0.839	0.942	0.913	0.927
	A_F_T	0.906	0.811	0.856	0.941	0.913	0.927
	A_L_F	0.939	0.760	0.840	0.938	0.909	0.923
	A_L_T	0.915	0.775	0.839	0.942	0.913	0.927
	L_T_F	0.927	0.777	0.845	0.941	0.908	0.924
Olympics	Baseline	0.969	0.455	0.619	0.881	0.823	0.851
	A_L_F_T	0.731	0.535	0.618	0.880	0.822	0.852
	A_F_T	0.778	0.446	0.567	0.881	0.817	0.848
	A_L_F	0.885	0.492	0.633	0.889	0.816	0.851
	A_L_T	0.742	0.506	0.602	0.888	0.819	0.852
	L_T_F	0.788	0.486	0.601	0.890	0.836	0.862
Halloween	Baseline	0.803	0.830	0.816	0.893	0.896	0.894
	A_L_F_T	0.811	0.834	0.822	0.890	0.883	0.886
	A_F_T	0.796	0.824	0.810	0.885	0.887	0.886
	A_L_F	0.823	0.822	0.822	0.891	0.883	0.887
	A_L_T	0.819	0.830	0.824	0.889	0.881	0.885
	L_T_F	0.825	0.826	0.826	0.889	0.880	0.884
HSandy	Baseline	0.989	0.844	0.911	0.972	0.931	0.951
	A_L_F_T	0.987	0.863	0.921	0.969	0.938	0.953
	A_F_T	0.872	0.913	0.892	0.965	0.938	0.952
	A_L_F	0.901	0.870	0.885	0.968	0.936	0.952
	A_L_T	0.987	0.863	0.921	0.969	0.938	0.953
	L_T_F	0.989	0.854	0.917	0.967	0.939	0.953

Source: the author

statistically significant improvements for the A_F_T combination. The improvements range from 0.7 to 4.5 percentage points in specific situations.

Considering the results presented in Table A.6 (i.e. F-Measure metric), we observed improvements in almost all combinations against the baseline. However, these improvements were not statistically significant.

In Table A.7, we present the results for Precision metric. A_L_F and L_T_F, were the combinations which resulted in less statistically inferior results, considering all datasets. On the other hand, the A_F_T was the combination with the greater number of statistically inferior results, which demonstrate the importance of the location core feature to achieve good results, considering this metric. The Halloween dataset was the only one that presented no statistically inferior results.

Table A.5: Results of statistical analysis, by comparing the Recall metric for each combination against the baseline

Dataset	Baseline	A_F_T	A_L_F_T	A_L_F	A_L_T	L_T_F
FaCup	0.820	**0.863 v**	0.846	0.837	0.846	0.846
Olympics	0.638	0.635	0.683	0.660	0.668	0.661
Halloween	0.858	0.854	0.857	0.852	0.856	0.855
HSandy	0.888	**0.924 v**	0.899	0.901	0.899	0.895

Source: the author

Table A.6: Results of statistical analysis, by comparing the F-Measure metric for each combination against the baseline

Dataset	Baseline	A_F_T	A_L_F_T	A_L_F	A_L_T	L_T_F
FaCup	0.878	0.892	0.884	0.884	0.884	0.888
Olympics	0.734	0.709	0.737	0.744	0.728	0.730
Halloween	0.854	0.848	0.854	0.854	0.855	0.857
HSandy	0.932	0.923	0.937	0.918	0.937	0.934

Source: the author

Table A.7: Results of statistical analysis, by comparing the Precision metric for each combination against the baseline

Dataset	Baseline	A_F_T	A_L_F_T	A_L_F	A_L_T	L_T_F
FaCup	0.954	0.925 *	0.929 *	0.941	0.929 *	0.936
Olympic	0.927	0.832 *	0.807 *	0.885	0.812 *	0.837 *
Halloween	0.851	0.843	0.852	0.858	0.856	0.860
HSandy	0.981	0.922 *	0.979	0.936 *	0.979	0.980

Source: the author

A.5 Conclusion

These preliminary experiments showed that the enrichment of specific features improved the results for planned and sportive events, mainly for the Recall metric. However, the improvements were modest, often at the expense of Precision. It also revealed that the core features defined are not strictly related to the event type.

We did not observe any representative pattern when considering the nature of the event, the number of conceptual features extracted from the datasets, and the presence or absence of specific features. Thus, this comparison enabled us to realize that no single combination of core features provides best results for all kinds of events. Based on these results, we decided to combine the core features according to the technique used to extract them. Thus, agents and locations are combined, producing the NER type of feature, and the frequent and representative terms are combined in the TERMS type of feature.

We also concluded that not all supplementary information incorporated to the datasets were discriminative enough to contextualize the tweets and improve significantly the classification performance. Considering the tweets as a whole, it is composed of poor textual content,

which may not provide all the information needed to characterize an event. As an attempt to overcome this problem, other sources of information related to the event will be explored.

Regarding the Semantic Enrichment step, the DBpedia knowledge base and the *rdf:type* property presented good coverage. However, it resulted in a huge amount of semantic features, some of them not discriminative to the event analyzed. Furthermore, only the application of the feature selection algorithm was not enough to select the most relevant textual and semantic features to improve the classification performance. Then, specific techniques to selected the most relevant semantic features according to the domain of the event need to be proposed.

A.6 Final Remarks

In this Appendix, we presented the experimental setup employed to analyze the contribution of each core feature for the Event Classification in tweets. The results show that each feature combination performs different, according to event analyzed. Furthermore, the presence or absence of specific features produced no expressive improvement or degradation on the results.

Nevertheless, the experiments defined key areas of the enrichment process to be improved, namely: a) use of external documents; and b) domain-specific pruning algorithm. It also showed that the core features could be grouped with no prejudice as entities and terms.

AppendixB

In this Appendix, we present the experiments in which we analyzed the performance of different thresholds proposed to be used in the Semantic Feature Pruning step. We also compared the influence of different feature selection techniques in these thresholds.

B.1 Pruning Thresholds

As mentioned in Chapter 4, we proposed a PageRank-based feature pruning algorithm to help in the selection of the most discriminative semantic features resulting from the Semantic Enrichment step. Different strategies were proposed to automatically define the pruning threshold, where the satisfactory ones were namely QUARTILES and IQR. The former produced the best results, such that it was selected to be used for the evaluation of the Hybrid Semantic Enrichment framework described in Chapter 5. The results of the IQR strategy are presented in this Appendix.

In summary, we aim at:

- comparing the results of both strategies (i.e. QUARTILES and IQR);
- analyzing the performance of the thresholds in combination with another feature selection technique;
- comparing the performance of the *CfsSubsetEval* and *Information Gain* algorithms available in Weka, as a complement to the feature selection process.

B.2 Experiment Description

In this section, we describe the experimental setup used to analyze the distinct thresholds proposed to be used in the Semantic Feature Pruning step and the influence of the feature selection techniques *CfsSubsetEval* and *InformationGain* in the Classification step. The same experiment configuration was employed to both strategies, following the setup described in Chapter 5:

- seven event target datasets, namely FaCup, HSandy, Halloween, Olympics, Alberta Floods, Australia Bushfire, and Influenza;

- the combination of textual and semantic features in the following datasets: NER (tweets uni-grams incorporated with the named entities semantically enriched); TERMS (tweets uni-grams incorporated with the frequent and representative terms semantically enriched); and ALL (tweets uni-grams incorporated with all conceptual features semantically enriched);
- baseline: composed by alphabetic uni-grams extracted from tweets.

To analyze the performance of the distinct thresholds definition strategies proposed for the pruning method, we prepared six (6) different setups for each semantically enriched dataset variation:

- using the QUARTILES threshold only;
- using the QUARTILES threshold in combination with the *CfsSubsetEval* algorithm;
- using the QUARTILES threshold in combination with the *Information Gain* algorithm;
- using the IQR threshold only;
- using the IQR threshold in combination with the *CfsSubsetEval* algorithm;
- using the IQR threshold in combination with the *Information Gain* algorithm.

In this Appendix, we report only the results for the IQR strategy. The results of the QUARTILES strategy are reported in Chapter 5 and will be used in the comparison section of this Appendix.

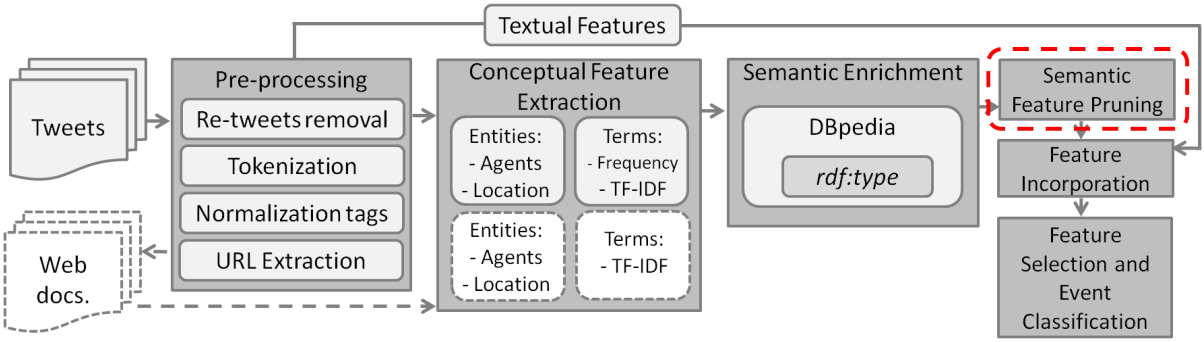
To classify the events, we used the algorithms NB and SMO with PolyKernel. We statistically compare the results using a two-tail paired $t - test$ from Weka Experimenter. We claim the improvement is *significant* with significance level of $\alpha = 0.05$.

B.3 Dataset Preparation

The focus of this analysis is the Semantic Feature Pruning step, as highlighted in Figure B.1. Thus, datasets were prepared exactly as described in Section 5.4. Pre-processing, Conceptual Feature Extraction, and Semantic Enrichment steps were executed following the *hybrid semantic enrichment* configuration described in Chapter 5.

In summary, we produced the following datasets:

Figure B.1: Summarized pipeline of the Event Classification process



Source: the author.

- semantically enriched datasets without the application of any pruning or feature selection algorithm (WP);
- IQR only;
- IQR in combination with *CfsSubsetEval* algorithm (i.e. IQR+CFS);
- IQR in combination with *Information Gain* algorithm (i.e. IQR+InfoGain).

B.4 Results and Discussion

In this Section, we analyze the performance of each Semantic Feature Pruning strategy, considering all target events, using IQR Strategy.

B.4.1 Qualitative and Quantitative Analysis of Selected Features

The same analysis was performed to the IQR strategy. In Table B.1, we present the number of textual features (TF), the amount of features resulting from the incorporation of textual and semantic features (WP), the number of features resulting from the Semantic Feature Pruning step (IQR), and its combination with the *Information Gain* (IQR+InfoGain) and *CfsSubsetEval* (IQR+CFS) algorithms. These results correspond to the ALL dataset of each configuration.

Considering the maximum value as the superior threshold, we amplified the range of semantic features that could be selected. Analyzing the textual and semantic features resulting from the Semantic Feature Pruning and Feature Selection step, we could observe that the characteristics of the algorithm applied for feature selection were of great importance, selecting more or less semantic features, and arranging them in different positions according to the

Table B.1: Resulting achieved by the IQR strategy

Dataset	TF	WP	IQR	IQR+InfoGain	IQR+CFS
FaCup	1672	2182	1978	1087	68
Olympics	1825	3723	3008	1260	90
Halloween	1829	4197	3307	1457	131
HSandy	2127	4311	3516	1630	70
Alberta Flood	1956	5068	3920	1627	38
Australia Bushfire	2092	4055	3303	1348	55
Influenza	1900	2657	2364	1123	51

Source: the author.

criteria used by the algorithm.

We analyzed the ALL datasets of each target event, considering both algorithms *CfsSubsetEval* and *InformationGain*. For the FaCup dataset, the semantic features are very representative, considering the *Information Gain* algorithm. This set of feature are extremely related to the domain of the event analyzed (e.g. *ontology/SoccerClub*, *yago/FormerFootballLeagueClubs*, and *ontology/SportsTeam*). The textual features that appeared soon after these semantic features were also linked to the topic analyzed (e.g. *#lfc*, *#cfc*). Considering the *CfsSubsetEval* algorithm, just a few semantic features appeared in the resulting list, but these features are very representative of the domain (e.g. *page/Category:Football_clubs_in_England*).

In the Olympics dataset, for both algorithms, most of the semantic features selected are related to location and the textual features refer to the topic of the event analyzed. The same pattern could be noticed in the results of the application of the *CfsSubsetEval* algorithm in the HSandy dataset. Considering the *InformationGain*, 28% of the resulting features are semantic features, most of them related to the event target.

Analyzing the Halloween dataset, we observed that the most relevant feature, after the *InformationGain* application, is the *T_URL* tag, following this textual feature, we have the semantic features that are extremely related to the domain of the event analyzed, such as *yago/TheSimpsonsCharacters* and *yago/FictionalCharactersIntroducedIn1987*. Considering the application of the *CfsSubsetEval* algorithm, the resulting list presented semantic features related to location and textual features related to the topic of the event.

For the Alberta Flood dataset, we observed a diversified set of resulting features. 41% of the resulting list is composed of semantic features. Most of them appeared in the top of the list refer to organization and location, such as *yago:ComputerSecurityOrganizations*, *yago/CharitableOrganizations*, and *ProvincesAndTerritoriesOfCanada*. Considering the dataset using the *CfsSubsetEval* algorithm, most of the semantic features were not directly related to the topic of the event analyzed.

For the Australia Bushfire dataset, discriminative semantic features were identified when using the *CfsSubsetEval* algorithm (e.g. yago:FireDepartment108121117). Applying the *Information Gain* algorithm, the semantic features were more related to organizations linked to the topic of the event analyzed (e.g. *yago/Service100577525*, *yago/EmergencyServicesInAustralia*).

In the Influenza dataset, as expected, the textual feature with the greater relevance is *flu*, after that, the semantic features presented in the resulting list refer to the location. Applying the *CfsSubsetEval* algorithm, the semantic features appear in the top of the list are also related to location entities.

B.4.2 Comparative Performance

After applying the NB and SMO algorithms: a) we analyzed the improvement achieved when using the pruning technique with the IQR thresholds (i.e. IQR only) against the same datasets, in which the Semantic Enrichment step was executed, but no pruning technique were used; and b) assessed the statistical significance of improvements achieved in the datasets using the IQR strategies (i.e. IQR, IQR+CFS, and IQR+InfoGain) against the textual baseline.

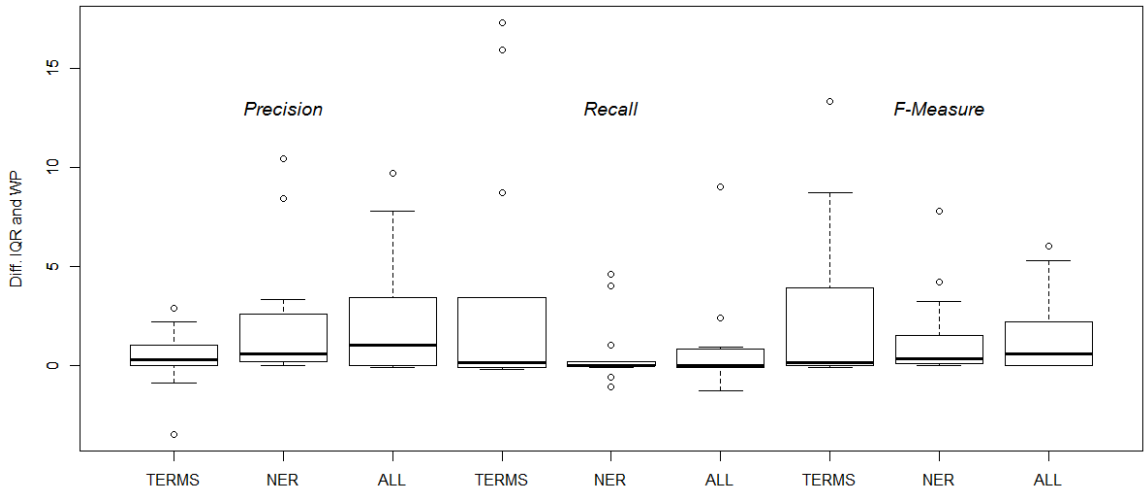
Figure B.2 presents the boxplot resulting from our first analysis considering the Precision, Recall, and F-Measure metrics for all types of features used (i.e. TERMS, NER, and ALL) and target events.

By applying the IQR strategy as the threshold, we were able to improve the results in 63.5% of the cases. All the types of features presented a small dispersion, specifically for Recall metric. Considering all the improvements obtained for each dataset and classification algorithm, the improvements achieved in the TERMS type of features were more expressive, adding a total of 49.3 pp.

Next, in the second comparison, we performed a statistical test in the results of the IQR strategy. Tables B.2, B.3, and B.4, summarize the results highlighting the ones that present statistically significant difference against the baseline, using the symbols (*) and (v) adopted in Chapter 5.

Table B.2 presents the results of the application of the IQR strategy, without any other feature selection technique. We can observe that a few results presented statistically significant difference against the baseline (i.e. 17%), most of them related to the Alberta Flood datasets. In general, we were able to improve the results in 37% of the cases, in which the improvements range from 0.1 pp to 7.9 pp for specific cases.

Figure B.2: Difference between using the hybrid semantic enrichment strategy in combination with IQR strategy and the baseline, in which no pruning and feature selection techniques was applied



Source: the author.

In Table B.3, we present the results of the application of the IQR strategy in combination with the *CfsSubsetEval* feature selection technique. To analyze this strategy, we also applied the *CfsSubsetEval* algorithm to the baseline. The combination of both strategies was able to statistically outperform the baseline in 54% of the case, of which 23 are related to the Recall metric (i.e. 18%). Most of the improvements could be noticed for the FaCup, Alberta Flood and Australia Bushfire datasets. Considering all the results, the improvements range from 0.1 pp to 8.8 pp in specific cases (i.e. for the Olympics dataset using the NER combination of features, considering the Recall metric).

The results for the application of the *InformationGain* algorithm in combination with the IQR strategy are presented in Table B.4. This combination was able to statistically outperform the baseline in 33% of the cases. The improvements range from 0.1 pp to 10.6 pp.

In general, the IQR+CFS combination presented the best results, outperforming the baseline in 54% of the cases. The improvements were very similar to the ones achieved in the QUARTILES+CFS strategy, discussed in Section 5.4.2. However, in the latter, the maximum value was much more expressive than the ones achieved with the IQR+CFS strategy. Considering the other combinations, without feature selection and using the *InformationGain* algorithm, the results achieved with the QUARTILES strategy presented better results when compared to the IQR strategy.

Table B.2: Statistical comparison between the baseline and the IQR strategy

IQR - without feature selection													
Dataset	Algor.	Baseline			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.936	0.809	0.867	**0.965 v**	0.590 *	0.732 *	0.952	0.653 *	0.774 *	0.945	0.738 *	0.828 *
	SMO	0.940	0.909	0.924	0.946	0.913	0.929	0.942	0.911	0.926	0.941	0.915	0.928
Olympics	NB	0.724	0.713	0.717	**0.803 v**	0.279 *	0.412 *	0.518 *	**0.787 v**	0.624 *	0.641 *	0.702	0.669 *
	SMO	0.885	0.823	0.853	0.880	0.823	0.850	0.878	0.828	0.851	0.883	0.833	0.857
Halloween	NB	0.859	0.733	0.790	0.813 *	0.757	0.783	0.662 *	0.692 *	0.676 *	0.680 *	0.693	0.686 *
	SMO	0.896	0.888	0.892	0.891	0.886	0.888	0.888	0.884	0.886	0.889	0.885	0.887
HSandy	NB	0.917	0.848	0.881	0.873 *	0.682 *	0.765 *	0.798 *	0.760 *	0.778 *	0.840 *	0.793 *	0.815 *
	SMO	0.966	0.919	0.942	0.969	0.923	0.945	0.955	0.920	0.937	0.955	0.920	0.937
Alberta Flood	NB	0.945	0.952	0.948	**0.967 v**	0.720 *	0.825 *	**0.996 v**	**0.994 v**	**0.995 v**	**0.996 v**	**0.993 v**	**0.995 v**
	SMO	0.999	0.992	0.995	0.996	0.989	0.993	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
Australia Bushfire	NB	0.931	0.960	0.945	0.937	0.742 *	0.827 *	**0.977 v**	**0.991 v**	**0.984 v**	**0.977 v**	**0.992 v**	**0.984 v**
	SMO	0.999	0.992	0.995	0.997	0.992	0.994	0.996	0.996	0.996	0.996	0.997	0.997
Influenza	NB	0.961	0.998	0.979	**0.986 v**	0.942 *	0.963 *	0.918 *	0.983 *	0.949 *	**0.981 v**	0.925 *	0.952 *
	SMO	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999

Source: the author.

Table B.3: Statistical comparison between the baseline and the IQR strategy in combination with the *CfsSubsetEval* algorithm

IQR+CFS													
Dataset	Algor.	Baseline			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.967	0.730	0.832	0.949 *	**0.796 v**	**0.865 v**	0.930 *	**0.789 v**	0.853	0.928 *	**0.803 v**	**0.861 v**
	SMO	0.978	0.769	0.861	0.968	**0.815 v**	**0.885 v**	0.966	**0.816 v**	**0.885 v**	0.968	**0.814 v**	**0.884 v**
Olympics	NB	0.970	0.454	0.616	0.848 *	0.389 *	0.533 *	0.934 *	**0.542 v**	**0.685 v**	0.814 *	0.489	0.610
	SMO	0.956	0.610	0.744	0.922 *	0.640	0.755	0.945	0.662	**0.778 v**	0.936	**0.657 v**	0.772
Halloween	NB	0.805	0.827	0.816	0.804	0.833	0.818	0.803	**0.867 v**	0.833	0.808	**0.867 v**	0.836
	SMO	0.840	0.877	0.858	0.843	0.881	0.861	0.847	0.874	0.860	0.846	0.879	0.862
HSandy	NB	0.976	0.822	0.892	0.884 *	**0.880 v**	0.882	0.922 *	0.843	0.881	0.904 *	**0.860 v**	0.881
	SMO	0.977	0.873	0.922	0.915 *	**0.907 v**	0.911	0.970	0.832 *	0.896 *	0.955 *	0.865	0.907
Alberta Flood	NB	0.991	0.971	0.981	0.981	**0.984 v**	0.983	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
	SMO	0.999	0.987	0.993	0.990 *	0.987	0.988	0.998	**1.000 v**	**0.999 v**	0.998	**1.000 v**	**0.999 v**
Australia Bushfire	NB	0.957	0.944	0.950	0.956	**0.977 v**	0.966	**0.982 v**	**0.999 v**	**0.990 v**	**0.982 v**	**0.999 v**	**0.990 v**
	SMO	0.992	0.981	0.986	0.979	0.980	0.980	0.997	**0.996 v**	**0.997 v**	0.998	**0.997 v**	**0.997 v**
Influenza	NB	0.999	0.997	0.998	0.999	0.998	0.998	0.999	0.996	0.997	0.999	0.995	0.997
	SMO	1.000	0.997	0.998	1.000	0.997	0.998	1.000	0.997	0.998	1.000	0.997	0.999

Source: the author.

Table B.4: Statistical comparison between the baseline and the IQR strategy in combination with the *InformationGain* algorithm

IQR+InfoGain													
Dataset	Algor.	Baseline			TERMS			NER			ALL		
		P	R	F	P	R	F	P	R	F	P	R	F
FaCup	NB	0.937	0.814	0.871	**0.964 v**	0.590 *	0.731 *	0.952	0.652 *	0.774 *	0.945	0.738 *	0.828 *
	SMO	0.972	0.910	0.940	0.975	0.910	0.941	0.971	0.914	0.941	0.972	0.914	0.942
Olympics	NB	0.769	0.691	0.727	0.805	0.272 *	0.405 *	0.534 *	**0.797 v**	0.639 *	0.659 *	0.703	0.679 *
	SMO	0.942	0.832	0.883	0.935	0.832	0.880	0.934	0.840	0.884	0.929	0.843	0.883
Halloween	NB	0.859	0.754	0.803	0.815 *	0.769	0.791	0.670 *	0.699 *	0.684 *	0.685 *	0.698 *	0.691 *
	SMO	0.921	0.904	0.912	0.919	0.906	0.912	0.915	0.904	0.909	0.916	0.904	0.910
HSandy	NB	0.927	0.854	0.889	0.872 *	0.681 *	0.764 *	0.799 *	0.771 *	0.784 *	0.838 *	0.797 *	0.817 *
	SMO	0.977	0.923	0.949	0.977	0.933	0.954	0.972	0.928	0.949	0.969	0.926	0.947
Alberta Flood	NB	0.955	0.954	0.954	0.968	0.722 *	0.826 *	**0.996 v**	**0.994 v**	**0.995 v**	**0.996 v**	**0.994 v**	**0.995 v**
	SMO	1.000	0.998	0.999	0.998	0.997	0.997	0.998	1.000	0.999	0.998	1.000	0.999
Australia Bushfire	NB	0.937	0.960	0.948	0.938	0.732 *	0.821 *	**0.976 v**	**0.991 v**	**0.983 v**	**0.977 v**	**0.992 v**	**0.985 v**
	SMO	0.996	0.995	0.996	0.997	0.994	0.996	0.995	0.996	0.996	0.995	0.997	0.996
Influenza	NB	0.967	0.998	0.982	**0.988 v**	0.939 *	0.963 *	0.928 *	0.982 *	0.954 *	**0.984 v**	0.923 *	0.953 *
	SMO	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999	1.000	0.997	0.999

Source: the author.

B.5 Performance Comparison Between the Strategies

In summary, the application of the Semantic Feature Pruning in combination with feature selection techniques have shown its striking role for improving the results, as well as in

Table B.5: Summary of the statistical test

Threshold	Without F.S.		CFS		InfoGain	
	v	*	v	*	v	*
IQR	22	38	40	17	16	39
QUARTILES	27	14	32	16	20	14

Source: the author.

the selection of the textual and semantic features related to the domain of the event analyzed. According to the technique selected, these related features can present more or less representativeness in the Classification step.

Considering the results presented in Tables B.2, B.3, and B.4, we observed that using only the Semantic Feature pruning technique, we were able to achieve good results and outperform the baseline in several cases. However, in combination with other feature selection techniques, particularly the *CfsSubsetEval*, the results are even better.

Comparing the strategies proposed for automatically defining the pruning thresholds, the result were very similar, with variations in the number of statistically significant results and the range of improvement achieved in each dataset, as presented in Table B.5. However, considering the number of cases in which the results were statistically inferior, the QUARTILES threshold faced this situation less often.

Considering the number of semantic features resulting from each threshold, using the QUARTILE strategy, we were able to select 30% less semantic features than the selected ones by the IQR strategy. Regarding the feature selection technique, the application of the *CfsSubsetEval* algorithm presented better results than the *InformationGain* algorithm.

Considering all these characteristics, the QUARTILES in combination with the *CfsSubsetEval* algorithm produced the better results in our analysis.

B.6 Final Remarks

In this Appendix, we presented the experimental setup employed to compare different thresholds for the Semantic Feature Pruning step, in combination with different feature selection techniques. The results show that the QUARTILES+CFS combination produced the better results.

AppendixC RESUMO EXPANDIDO

Um Framework para Classificação de Eventos em Tweets Baseado em Enriquecimento Semântico Híbrido

Plataformas de mídias sociais são amplamente utilizadas para o compartilhamento de informações sobre assuntos variados ao redor mundo. Dentre elas, o Twitter acabou se tornando uma importante fonte de dados em diversas aplicações, devido ao grande volume de mensagens compartilhadas todos os dias e a grande variedade de assuntos abordados nessas mensagens.

Dentre essas aplicações, podemos mencionar a análise de sentimento, mineração de opinião, detecção de eventos, identificação de notícias de última hora, entre outras. Em relação à detecção de eventos em *tweets*, ainda não há um consenso quanto a quais são as tarefas que compõem a área de Identificação e Classificação de Eventos. No geral, a tarefa de *identificação de eventos* está relacionada à criação de conjuntos de mensagens de acordo com o tópico que abordam ou com o período em que foram publicadas. Já a tarefa de *classificação de eventos* visa construir modelos para filtrar e categorizar essas mensagens. Neste contexto, este trabalho visa utilizar as postagens compartilhadas no Twitter como base para a classificação de eventos.

Contudo, a classificação de eventos em tweets é uma tarefa não trivial, que exige mais do que a aplicação de técnicas convencionais de Processamento de Linguagem Natural (PLN) e abordagens para classificação de texto. Isso é devido, principalmente, à dinamicidade da linguagem e ao vocabulário pobre utilizado pelos usuários desta plataforma.

Como uma tentativa de amenizar esses problemas, diversas abordagens foram propostas para agregar contexto externo a essas mensagens. Dentre as técnicas de enriquecimento mais utilizadas, podemos mencionar a utilização de: a) *documentos externos* relacionados ao evento, como uma forma de incorporar conteúdo do domínio através da análise de páginas web (e.g. blogs, sites e artigos da Wikipédia); b) *bases de conhecimento*, como as disponibilizadas pela Linked Open Data (LOD) cloud; e c) ferramentas para reconhecimento de *entidades nomeadas*, as quais podem ser utilizadas para a generalização de categorias específicas de entidades. É importante ressaltar que tais soluções podem acrescentar uma grande quantidade de novas *features*, muitas das quais não comtribuem efetivamente para a caracterização do evento analisado ou seu domínio.

Além disso, cada trabalho considera um tipo de evento diferente, tendo como base suposições específicas alinhadas ao objetivo da aplicação. Logo, técnicas de enriquecimento são utilizadas tendo como base diferentes critérios, bem como distintas *features* textuais e fontes de informação, tornando difícil a reprodução dessas abordagens em outros tipos de eventos, assim

como a comparação entre elas.

Nesse contexto, o objetivo deste trabalho é *"propor a construção de um framework para classificação de eventos em tweets, que possa ser reproduzido e comparado considerando eventos de natureza distinta, utilizando como base o enriquecimento semântico híbrido"*. Os objetivos específicos são: a) identificar as diferentes definições de evento e as *features* que são utilizadas para caracterizá-lo; b) identificar recursos externos de informação que possam ser utilizados para enriquecer o conteúdo dos tweets com informação contextual; c) definir um processo para enriquecer semanticamente o conteúdo dos tweets, de forma que isso possa ser aplicado a eventos de natureza distinta e assim contribuir para a melhoria da tarefa de classificação de eventos; d) desenvolver um conjunto de experimentos que nos auxiliem a mensurar a contribuição do enriquecimento semântico na classificação de eventos em *tweets*.

Para atingir este objetivos, são exploradas neste trabalho as seguintes questões de pesquisa: a) *existe um conjunto de features que é mais discriminante em determinados tipos de eventos?*; b) *qual é o processo de enriquecimento semântico, juntamente com recursos e propriedades, que podem trazer melhores resultados para a classificação de eventos em tweets, e como aplicá-lo?*

Quanto às definições de evento, na literatura foi possível identificar duas categorias principais para evento, sendo elas: eventos *especificados*, quando se sabe exatamente o evento que se pretende observar, e *não especificados*, quando se está, por exemplo, monitorando o compartilhamento de mensagens na plataforma, e se observa o aumento repentino de um determinado conjunto de termos, os quais podem descrever a ocorrência de um evento. Eventos especificados e não especificados podem ser divididos em eventos *planejados* e *não planejados*. Em eventos planejados, possuímos informações prévias sobre o evento, como por exemplo, o local onde ocorre e as bandas que se apresentarão em determinado festival musical. Já os eventos não planejados estão associados a incidentes e desastres naturais, como acidentes de carro e terremotos.

Além dessas categorias, um evento pode apresentar algumas características básicas, como: a) estar sempre associado a um tópico ou assunto, que geralmente é representado por um conjunto de termos chave; b) componente temporal, que representa o período no qual o evento ocorreu; c) escala, que representa a relevância e o impacto causado pelo evento; d) propriedades geográficas associadas ao local onde o evento ocorreu; e) agentes que estão direta ou indiretamente relacionados ao evento.

Com base nas definições apresentadas, neste trabalho foi proposto que: *"um evento é uma ocorrência representada por um tópico que acontece em um período específico e pode*

envolver um ou mais locais e agentes.". Dessa forma, o framework atende à classificação de eventos especificados e de qualquer natureza.

Em relação ao processo de enriquecimento, os trabalhos relacionados foram analisados quanto à forma como o enriquecimento contextual é realizado. Como foi mencionado, existem três técnicas principais, sendo a utilização de documentos externos, o enriquecimento semântico, e a utilização de ferramentas para o reconhecimento de entidades nomeadas. Esta última pode ser utilizada como uma ferramenta complementar em ambos os casos.

Para a análise, os trabalhos relacionados foram agrupados quanto ao tipo de enriquecimento aplicado. Temos então o enriquecimento externo, quando são extraídos destes documentos entidades nomeadas, *hashtags*, termos representativos utilizando TF-IDF, entre outros, para se identificar páginas *web* relacionadas, como *blogs*, sites de notícias ou artigos da Wikipédia. Estas diferentes informações extraídas são agregadas às caractrísticas extraídas dos próprios *tweets*. As limitações dessa abordagem são definir qual conteúdo extrair e quais dados são mais importantes para a classificação de eventos em *tweets*.

Outros trabalhos utilizam apenas o enriquecimento semântico, através da utilização de bases de conhecimento, como as disponibilizadas pela LOD *cloud*. Para a identificação dos conceitos na base de conhecimento são utilizadas entidades nomeadas, TF-IDF, expressões de tempo e espaço, entre outros. Tais conceitos são utilizados para generalizar o conteúdo dos *tweets*, através da agregação de informações/objetos de uma determinada propriedade. Os desafios desta abordagem são saber quais dados enriquecer, assim como qual base de conhecimento e propriedades explorar.

Por fim, temos trabalhos que combinam ambas as abordagens e utilizam *features* tanto dos *tweets* como de documentos e bases externas para auxiliar na identificação de conceitos semânticos e assim melhorar os resultados da tarefa de classificação de eventos. Tais trabalhos estão diretamente relacionados com a abordagem proposta nesse trabalho, cujo objetivo é utilizar o enriquecimento semântico híbrido.

Em resumo, as abordagens propostas diferem em relação ao tipo de evento a ser considerado na aplicação. Diferentes soluções são utilizadas para agregar contexto aos *tweets* e auxiliar na tarefa de classificação. Além disso, os trabalhos também diferem em relação às *features* que são utilizadas, uma vez que, elas são selecionadas conforme o objetivo da aplicação. Em resumo, as abordagens são específicas para cada aplicação, e nem sempre podem ser aplicadas em outras situações.

Com base nos trabalhos relacionados, propôs-se um framework para a classificação de eventos em *tweets*, cujas principais características são: a) um conjunto de *core features* que

melhor caracterizam um evento; b) um processo de enriquecimento híbrido, o qual é composto por enriquecimento tendo como base documentos externos, enriquecimento semântico e a utilização de ferramentas de reconhecimento de entidades nomeadas; c) um algoritmo para a seleção de *features* semânticas mais discriminativas, dado o domínio do evento analisado. Além disso, a abordagem visa ser generalizável, ou seja, aplicável a eventos de qualquer natureza sem seguir suposições específicas, de forma que ele possa ser reproduzido e comparável com outros eventos e outras abordagens.

Para responder então a primeira questão de pesquisa (*Existe um conjunto de* features *que é mais discriminante em determinados tipos de eventos?*), foi possível identificar um conjunto de *features* comum à maioria dos trabalhos: a) agente, que representa pessoas ou organizações que são direta ou indiretamente afetados pelo evento; b) local, que representa o local onde o evento ocorreu, local do agente que atuou ou reportou o evento; c) vocabulário do domínio, o qual pode ser representado por termos frequentes e termos representativos utilizando determinada técnica de pesagem.

Após uma série de testes e experimentos, para responder a segunda questão de pesquisa (*Qual é o processo de enriquecimento semântico, juntamente com recursos e propriedades, que podem trazer melhores resultados para a classificação de eventos em* tweets*, e como aplicá-lo?*), foi elaborada a estrutura para o framework, a qual é composta por seis etapas principais.

Na etapa *pre-processing*, dado um conjunto de *tweets*, é realizado um pré-processamento básico nos dados, como remoção de re-tweets, tokenização e normalização de símbolos específicos (i.e. *emoticons*, anotação de usuário e URLs), resultando em um conjunto de *features* textuais. Representando o enriquecimento através de documentos externos, dado como entrada as URLs que foram identificadas nos *tweets*, nesta etapa o conteúdo dessas URLs também é extraído e utilizado como entrada para a etapa seguinte.

Em seguida, na etapa *contextual feature extraction*, tendo como base os *tweets* pré-processados e o conteúdo extraído das páginas *web* relacionadas, é executada a extração de *features* conceituais, através de ferramentas de reconhecimento de entidade nomeada e outras ferramentas para a extração e contagem de termos. As *features* extraídas são aquelas mencionadas anteriormente, agentes, locais e termos do domínio. É importante ressaltar que o conjunto de *features* extraídas é diferente para cada documento, as quatro *core features* são extraídas dos *tweets*, mas apenas três delas são extraídas dos documentos externos (i.e. agentes, locais e termos representativos), uma vez que o texto dessas páginas, geralmente, possui muitas informações e experimentos preliminares mostraram que os termos resultantes da aplicação de TF-IDF (i.e. termos representativos) acabavam por englobar os termos frequentes. Logo, o

resultado desta etapa é um conjunto de *features* conceituais.

As *features* conceituais são então utilizadas como base para a etapa *semantic enrichment*, na qual podem ser utilizadas diferentes bases de conhecimento, como as disponibilizadas pela LOD *cloud* e diferentes propriedades podem ser exploradas. O resultado desta etapa é um conjunto de *features* semânticas, as quais nos auxiliarão a melhorar o desempenho da tarefa de classificação de eventos em *tweets*.

O processo de enriquecimento semântico pode retornar um grande volume de novas *features*, sendo que algumas delas podem não ser representativas para o domínio do evento sendo analisado. Desse modo, na etapa *semantic feture pruning* foi proposta a utilização de um método de poda para reduzir o número de *features* segundo determinado critério. O algoritmo de poda proposto é baseado no PageRank, o qual tem como função básica a atribuição de peso aos nós de um grafo de relacionamentos. Esse algoritmo é utilizado juntamente com um método para definição automática dos limiares que definem quais *features* são consideradas específicas ou genéricas de mais para o evento.

Em seguida, na etapa *incorporation*, as *features* semânticas resultantes da poda são incorporadas ao conjunto de *features* textuais extraídas do *dataset* alvo. Passada essa etapa, o *dataset* enriquecido está pronto para a tarefa de classificação e, se necessário, algoritmos para seleção de atributos de propósito geral podem ser aplicados (i.e. etapa *feature selection and event classification*).

Definida então a estrutura do framework, foram elaborados dois experimentos principais para validá-lo. O primeiro experimento tem por objetivo verificar a contribuição do algoritmo de poda proposto, assim como da utilização de técnicas de propósito geral para a seleção de atributos, além de comparar o desempenho do enriquecimento híbrido com a utilização de enriquecimento semântico apenas. Dado que encontramos a configuração que produz os melhores resultados, o segundo experimento visa comparar o desempenho do enriquecimento híbrido com a utilização de um método alternativo para a execução do enriquecimento contextual, chamado *Word Embeddings*.

Para a realização dos experimentos, foram utilizados sete *datasets* alvo, os quais representam eventos de natureza distinta (e.g. eventos esportivos, datas comemorativas, desastres naturais e epidemias). Os datasets representam eventos distintos para testar a capacidade de generalização do framework, ou seja, verificar se ele é capaz de apresentar bons resultados para qualquer tipo de evento. Tendo como base experimentos anteriores e a forma como as *features* são extraídas, optamos por agrupar as entidades nomeadas (i.e. agentes e locais) em um conjunto denominado NER, e os termos frequentes e termos representativos em outro conjunto

chamado TERMS. Para a realização do enriquecimento semântico, DBpedia foi escolhida por ser uma base de conhecimento *cross-domain* com informações sobre diversos assuntos. Dentre as propriedades disponibilizadas por essa base, foi escolhida a propriedade *rdf:type*. Para a definição automática dos limiares de corte, foram testadas duas estratégias principais, sendo que QUARTILES foi a escolhida, devido aos ótimos resultados apresentados.

Como o foco dos experimentos é diferente, dois *baselines* distintos foram utilizados na comparação. O primeiro é composto apenas por *features* textuais, para o experimento em que desejamos comparar a contribuição do enriquecimento semântico. No segundo *baseline*, o enriquecimento dos *datasets* foi realizado através da abordagem de Word embeddings, utilizando vetores de palavras de 100 dimensões. Para cada evento alvo foram criados três *datasets*, cada um contendo *features* enriquecidas conforme os três conjuntos de *features* definido: enriquecendo apenas entidades nomeadas (NER), apenas os termos que representam o vocabulário do domínio (TERMS) e a junção de ambos (ALL). Foram testadas diferentes configurações, nas quais foram feitas variações quanto a forma como são selecionadas as *features* mais discriminativas, assim como na forma como o enriquecimento contextual é realizado (i.e. utilizando documentos externos ou não). Todos os *datasets* foram submetidos aos classificadores Naive Bayes e SMO, no ambiente Weka. E para a análise dos resultados foi executado o teste estatístico *teste-t* pareado, considerando um nível de confiança de 95%.

O experimento 1 foi divido em três para facilitar a análise. O experimento 1.1 tem por objetivo analisar o desempenho do método de poda, logo para esse experimento, não foi aplicado nenhum algoritmo para a seleção de atributos de propósito geral. A configuração produzida neste experimento recebe o nome de QUARTILES, que representa a técnica para definição automática dos limiares de poda utilizada. O experimento 1.2 tem por objetivo avaliar a desempenho do framework utilizando somente algoritmos para a seleção de atributos de propósito geral, assim como verificar se a combinação desses algoritmos com o método de poda proposto neste trabalho é capaz de melhorar ainda mais os resultados da tarefa de classificação. Para o experimento 1.2, temos as seguintes configurações: a) CFS, para a aplicação do algoritmo *CfsSubsetEval*; b), InfoGain, para o a aplicação do algoritmo *Information Gain*; c) QUARTILES+CFS, para a combinação do método de poda proposto neste trabalho com o algoritmo de propósito geral *CfsSubsetEval*; e) QUARTILES+IG, para a combinação do método de poda proposto neste trabalho com o algoritmo de propósito geral *Information Gain*.

Primeiramente, os resultados foram analisados em relação ao volume de *features* que cada configuração foi capaz de reduzir, além disso, foi analisado manualmente a representatividade desses *features* para o evento sendo analisado. Considerando o volume de *features*

textuais e semânticas resultantes após a aplicação do processo de enriquecimento semântico e incorporação, para a configuração QUARTILES, em média, tivemos uma redução de 41% no volume total de *features*. Aplicando o algoritmo de seleção de atributos *Information Gain*, o volume de *features* foi reduzido em 56,59%, em média. Os resultados mais significativos foram encontrados para a configuração CFS, a qual foi capaz de reduzir o volume total de *features* em 97,94%, em média.

Em uma análise manual, foi possível observar que o algoritmo de poda proposto neste trabalho foi capaz de selecionar *features* que são representativas para domínio do evento analisado, porém como o algoritmo atua somente sob as *features* semânticas essa redução não foi muito acentuada. Quanto aos algoritmos de propósito geral, CFS e InfoGain possuem critérios distintos, mas em ambos os casos, as *features* semânticas sempre foram representadas no topo da lista, demonstrando sua importância para a tarefa de classificação. No geral, foi possível observar que do total de *features* resultantes após a combinação de ambos os métodos, em média 30% corresponde a *features* semânticas.

Analisando então o impacto dessa redução no número de *features* na tarefa de classificação, temos os seguintes resultados: a) a configuração QUARTILES+CFS apresentou melhorias em mais casos (53,1%); b) as melhorias foram mais significativas, considerando a aplicação do $test - t$, usando apenas CFS (28,5%); c) a configuração QUARTILES+CFS produziu melhorias de até 32,6 pontos percentuais em comparação com o *baseline* textual, sendo que tal melhoria é quatro vezes superior aos resultados encontrados com a configuração CFS. Com relação aos tipos de *features* que obtiveram melhor resultado, entidades nomeadas aparecem em todos os casos, seja isolado, como nos *datasets* NER ou em combinação com o vocabulário do domínio, neste caso, ALL *datasets*.

Recall foi a métrica que apresentou uma quantidade maior de casos com melhorias. Quanto aos *datasets*, melhorias acentuadas foram encontradas em eventos que mencionam desastres naturais e eventos esportivos. No geral, resultados obtidos com CFS e sua combinação com o algoritmo de poda foram os resultados mais promissores.

Dado então que avaliamos a contribuição dos algoritmos para seleção de atributos, seja de propósito específico ou geral. No experimento 1.3 analisamos a contribuição do enriquecimento utilizando documentos externos para a tarefa de classificação de *tweets*. Para tanto, duas configurações foram consideradas, a aplicação do enriquecimento semântico apenas, no qual são utilizadas apenas as *features* conceituais extraídas dos *tweets*, e a aplicação do enriquecimento semântico híbrido, quando são utilizadas como base para o enriquecimento *features* provenientes dos *tweets* e dos documentos externos. Neste experimento, foram testados apenas

QUARTILES+CFS e CFS, pois foram as configurações que apresentaram melhores resultados nos experimentos anteriores.

Analisando a quantidade de *features* semânticas resultante em cada configuração, foi possível observar que o enriquecimento híbrido aumenta o número de *features*, em média, em 240%. Quanto ao impacto da adição dessa enorme quantidade de novas *features* no desempenho da tarefa de classificação, observamos que a utilização do algoritmo de seleção de *features* CFS apresenta bons resultados tanto para a versão com enriquecimento híbrido (52,3%), quanto para a utilização de enriquecimento semântico apenas (59,5%). Já para a aplicação do método de poda (i.e. QUARTILES+CFS), a configuração de enriquecimento híbrido apresenta resultados superiores (53,17%), principalmente em termos estatísticos (25,4%). Novamente, entidades nomeadas estão presentes na maioria dos casos que apresentaram os melhores resultados. Quanto à métrica que apresentou mais melhorias, Recall foi unanime, dando evidências de que o enriquecimento semântico foi capaz de generalizar o conteúdo dos *tweets* e assim auxiliar na recuperação de mais *tweets* relacionados ao mesmo evento.

No geral, considerando os três experimentos, temos que o enriquecimento híbrido é capaz de melhorar o desempenho da classificação de eventos em *tweets*. Entidades nomeadas estão presentes na grande maioria das configurações que apresentaram os melhores resultados. Recall foi a métrica com melhores resultados na maioria dos casos, dando evidências da capacidade de generalização do enriquecimento semântico. Para o enriquecimento híbrido, os melhores resultados foram encontrados combinando o algoritmo de poda proposto e o algoritmo de propósito geral CFS, gerando evidências de que o algoritmo de poda atuou como uma etapa de pré-filtragem, entregando para o algoritmo *CfsSubsetEval* as *features* semânticas mais representativas. Já para a utilização de enriquecimento semântico apenas, os melhores resultados foram encontrados com a utilização do algoritmo CFS apenas. Quanto a técnica de enriquecimento utilizada, enriquecimento externo melhorou os resultados, principalmente dos *datasets* nos quais muitas *features* foram incorporadas. Esse resultado levanta a hipótese de que quando se conhece de antemão o dataset, ou que para o tipo de evento em questão as URLs dificilmente estarão disponíveis, por exemplo, para um *dataset* muito antigo, o melhor é utilizar enriquecimento semântico apenas. Mas quando o *dataset* é mais recente, o melhor é aplicar o enriquecimento híbrido, pois as chances são maiores de que as URLs ainda estejam ativas. Os experimentos apresentaram bons resultados para tipos de eventos distintos, mostrando a habilidade do framework em lidar com eventos de diversas naturezas.

Uma vez encontrada uma configuração capaz de melhorar o desempenho da tarefa de classificação de eventos, foi decido então comparar o framework proposto com uma técnica

alternativa para a realização do enriquecimento contextual, chamada Word embeddings. O enriquecimento híbrido, com a utilização da técnica de poda e do algoritmo de seleção de *features* (i.e. QUARTILES+CFS) foi a configuração utilizada neste experimento.

Os *datasets* criados utilizando Word embeddings foram submetidos aos classificadores Naïve Bayes e SMO. Em comparação com os resultados obtidos pela técnica proposta neste trabalho, o framework foi capaz de melhorar os resultados em 95,2% dos casos. Para cada métrica, a média de melhoria foi de 12 pontos percentuais para F-Measure, 16 pontos percentuais para Precisão e 15 pontos percentuais para Recall. Considerando a análise estatística, em 83% dos casos, os resultados apresentaram diferença estatisticamente significativa.

Quanto às contribuições deste trabalho: a) foi identificado um conjunto de *features* capaz de representar os mais variados tipos de eventos; b) foi proposto um processo para o enriquecimento semântico híbrido, no qual essas *features* são extraídas tanto do conteúdo dos *tweets* como de documentos externos relacionados; c) foi proposto um método para a seleção das *features* semânticas mais discriminativas para um dado domínio; d) foram executados experimentos com sete *datasets* representando eventos distintos, para analisar a capacidade de generalização do framework; e) a abordagem proposta neste trabalho foi comparada com outra técnica para o enriquecimento contextual chamada Word embeddings; f) os resultados parciais deste trabalho foram apresentados em duas conferências.

Como conclusão: a) os melhores resultados foram encontrados com as configurações de *datasets* em que entidades nomeadas foram utilizadas; b) a utilização de documentos externos foi capaz de adicionar informações relevantes sobre o domínio do evento analisado e assim contribuir para a melhoria do desempenho da tarefa de classificação de eventos em tweets; c) o enriquecimento semântico utilizando a base de conhecimento DBpedia foi capaz de generalizar as informações sobre o evento com uma cobertura média de 80%; d) o método de poda proposto apresentou bons resultados para a redução das *features* semânticas, mas sua combinação com CFS apresentou melhorias mais significativas no desempenho da tarefa de classificação; e) os experimentos possibilitaram a verificação e confirmação de que o framework é uma solução generalizável que pode ser aplicada para a classificação de eventos de natureza distinta, sem seguir nenhuma suposição específica.

Quanto às limitações e trabalhos futuros: a) outras bases de conhecimento assim como propriedades disponibilizadas por essas bases podem ser empregadas; b) o método de seleção de *features* apresentou bons resultados, mas outros critérios podem ser utilizados; c) foram utilizados apenas sete *datasets*, o que pode ter dificultado a identificação e confirmação de alguns padrões, logo como trabalhos futuros temos a inclusão de novos *datasets*; d) utilização de out-

ros algoritmos de classificação que, segundo a literatura, tem apresentado bons resultados para esse tipo de tarefa; e) criação de um modelo capaz de classificar *tweets* relacionados ao mesmo evento, mas que ocorreram em períodos de tempo distintos; f) definição de uma arquitetura que além do processo de classificação, nos permita realizar também a tarefa de identificação de eventos em *tweets*; g) elaboração de um conjunto de experimentos que nos permita comparar essa abordagem com outras abordagens do estado da arte.