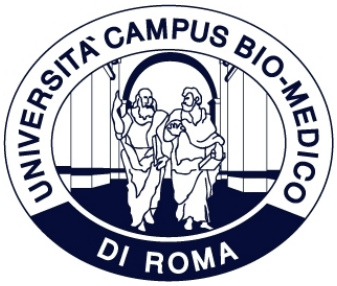


μLevel Rumour Detection on Twitter

Candidate: **Rosa Sicilia**

Supervisor: **Prof. Paolo Soda**

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Bioengineering and Bioscience

XXXII cycle

Academic Year 2016 - 2020

Computer Science and Bioinformatics Laboratory

Department of Engineering

Università Campus Bio-Medico di Roma

January, 2020

Rosa Sicilia

ABSTRACT

Recent years have witnessed a drastic chance in information diffusion that has become more and more immediate and effortless thanks to social media, allowing not only certified and professional press practitioners but also common end users to share news contents at a little to no cost. Despite the clear advantages of this phenomenon, the absence of systematic control and moderation on these platforms easily leads to spread unreliable information. This is usually referred to as rumour, an unverified and instrumentally relevant statement in circulation. To prevent treacherous information to have social consequences, researchers have been directing considerable effort in studying automatic systems able to recognize rumours. Most of the work focuses on macro-level analyses, i.e. the detection system considers as rumour news carried by a set of microblog posts rather than by an individual post. However, a micro-level analysis that considers the individual posts, could be of major interest in specific domains, such as health, where a finer investigation is often needed. On these grounds, in this thesis we investigate machine learning methods to detect rumours at the micro-level, and we apply our research on two real-world test cases. To this goal, we investigated four main directions: the collection and the annotation of two datasets, the design of the feature set, the introduction of a novel feature selection approach and the investigation of how the knowledge can be transferred among different topics. First, we present two Twitter datasets on health trending topics, manually labeled in three classes, namely rumour, non-rumour, i.e. referenced news, and unknowns, i.e. posts that do not belong to rumours and non-rumours classes. Second, we design a novel feature set, accounting both descriptors based on the literature and newly conceived for the micro-level task, describing influence potential and network characteristics. Third, we explore the feature selection influence on the specific problem, proposing a novel filter algorithm, relying on a rule-based topology framework which characterizes the feature space aiming at reducing samples in unreliable configurations. Testing this third approach on two health datasets, we are able to obtain promising results, reaching even an accuracy of 96.8%. As a further step in micro-level research, we also tackle the problem of knowledge transfer among different topic domains. To this end, we present a novel hybrid transfer learning approach that exploits the rule-based topology framework used for feature

Rosa Sicilia

selection. Comparing this novel method with sate-of-the-art techniques over our two datasets we are able to provide interesting results, showing the validity of the method and the potential of transfer learning for rumour detection.

Rosa Sicilia

TABLE OF CONTENTS

Rosa Sicilia

Rosa Sicilia

Rosa Sicilia

LIST OF FIGURES

Rosa Sicilia

Rosa Sicilia

Rosa Sicilia

LIST OF TABLES

Rosa Sicilia

Rosa Sicilia

Rosa Sicilia

CHAPTER 1

"Rumour has it.."

1.1 Social microblogs and misinformation

In recent years the generation and diffusion of information has significantly changed thanks to the rise of social media platforms, and in particular of social microblogs, which have emerged as the most used social network services. These resources provide the users a constantly updated pool of news that ranges over all possible subjects, and the user itself is able to share and propagate information immediately and at little to no cost. This empowerment concerns not only certified and professional press practitioners, but also common end-users, revealing an important difference between social microblogs and traditional news sources. In fact, despite the clear advantages due to the access to an unprecedented source of information, the absence of systematic control and moderation of the posts on these platforms easily leads to misinformation spread. This reality should not be underestimated since recent surveys conducted in the United States have shown that people rely more and more on the news found on social media (Perrin 2015, Shearer and Gottfried 2017). Hence, it is of paramount importance that such news that are unverified or even false are effectively monitored and debunked. Such unreliable information in terms of unverified and instrumentally relevant statements in circulation is usually referred to as *rumour* (DiFonzo and Bordia 2007). According to the literature, this term often denotes information which is ultimately deemed false (Zubiaga et al. 2016, 2018): this definition however leaves out both the newsworthy stories and the potential threats that, having the same level of uncertainty at a first time spread, only in the end reveal themselves to be true. For this reason we prefer to adopt a more general and comprehensive approach defining a rumour as unverified news in circulation with an instrumental value and likely to be dangerous (DiFonzo and Bordia 2007).

Social media can be considered the primary vehicle of such treacherous information diffusion, thanks to the possibility to easily share and widely spread news in short time. For instance, in 2013 the official Twitter account of the Associated Press was hacked and it sent out a rumour about the explosion of two bombs at the White House and the US President being injured in the attack. This news caused deep panic in such a broad scale that ended with a dramatic, though brief, crash of the stock market RumorWhiteHouse.

Rosa Sicilia

Another tragic episode happened in 2015, when unverified news about shoot-outs and kidnappings in the schools of Veracruz spread on Twitter and Facebook, causing chaos in the city and involving 26 car crashes (Ma et al. 2016). Along with these events, other similar episodes of misinformation causing disasters can be found in many domains, such as emergency situation management, health and wellness, politics, etc. (Guan et al. 2014, Zubiaga et al. 2016, Chung and Zhang 2017).

On these motivations, researchers have been directing considerable efforts towards the study of this phenomenon and of the means to handle it.

1.2 Tackling rumour diffusion

Traditionally, rumours spread by means of word of mouth, radio or newspapers, but the recent changes in information diffusion, turned into changes in the research fields, leading from the social and psychological point of view to the scientific computational analysis. In other words the unprecedented amount of data generated by social microblogs day by day, has straightforwardly led to the need of realizing automatic processes able to detect such rumours.

Most of the current literature focuses on detecting rumour information at an aggregated level, hereinafter also referred to as *macro-level*: this means that the detection system considers as rumour news carried by a set of microblog posts rather than by an individual post. These sets could be related to the same conversational thread (Wu et al. 2015), to a specific topic (Castillo et al. 2011, Kwon et al. 2013, Ma et al. 2015), aggregated according to a particular event (Ma et al. 2016, Kwon et al. 2017, Wang, Guo, Wang, Li and Tang 2019) or clustered according to the content and enquiry level (Zhao et al. 2015).

However, being able to detect rumour at the *micro-level* of a single post rather than at the macro-level can be very, and even more, useful in many applications. Indeed, for example in security issues it could be necessary to immediately know the exact source of a rumour, this would need a further analysis of the single posts of a rumour found with a macro-level approach. Moreover, the literature has shown that rumours concentrate in some specific domains (Wu et al. 2015) such as health care. In this domain a micro-level analysis could be more efficient: in fact, it is often possible to find rumour and non-rumour information belonging to the same conversation or topic. In addition, nowadays people often look for health knowledge and advices on online services, but not all these resources provide accurate or reliable information (Wang, McKee, Torbica and Stuckler 2019). Hence, information management in this particular field could have a great impact improving life quality and style of those people that rely on health knowledge found online, especially in social microblogs, which allow a faster and broader diffusion of news.

Rosa Sicilia

For this reason we deem that it is crucial to conduct a finer analysis, detecting rumour at the *micro-level* of a single posts and, in this respect, there is a large need to investigate this topic. Indeed, to the best of our knowledge, further to this work, only the work by Zubiaga et al. (2017) and the most recent work by Fard et al. (2019) tackle this issue. However the former only considers those news that generated as high number of sharings, leaving out all those small conversations or isolated posts that could be the start if a rising rumour, whereas the latter introduces a one-class classification approach.

1.3 Contributions

For all the aforementioned reasons, the main focus of this work is to advance in the underrepresented research on micro-level rumour detection and it presents a novel machine learning-based system focusing on topic-specific rumours on Twitter social microblog. We address the problem under different perspectives, starting from the features design, analysing the feature selection process and finally trying to transfer knowledge between different topic domains.

In detail, in chapter 4 we explore two levels of features, named as user level and network level, respectively. For each level, we study three types of features including influence potential, personal interest and network characteristics. In particular, we also develop new descriptors including the likelihood that a tweet is retweeted and the likelihood of a URL to be shared, conversation size, fraction of users followers of root, and fraction of tweets with URLs in a conversation. We investigate the discrimination power of different features using different classifiers. To prove that the use of general features unrelated to the particular topic allows to employ the developed approach on different specific domains, we studied the robustness of the features when they are applied on different topic-specific datasets.

In chapter 5 we introduce a novel method for feature selection relying on a rule-based topology framework which characterizes the feature space aiming at reducing samples in unreliable configurations. Moreover, the proposed method is compared with current state-of-the-art techniques in order to asses its validity and show its potential.

Finally, in chapter 6 we explore knowledge transfer between different topics. We start investigating how the change of topic can affect the performance of a rumour detection system in cross-topic tests, i.e. when the system is trained on posts with a topic and tested on samples with a different one. Then we explore current state-of-the-art instance-based and feature-based transfer learning methods, presenting a novel hybrid solution that exploits both transfer types. This is based also on the rule-based topology framework presented for feature selection, showing its adaptability to different applicative contexts.

Rosa Sicilia

CHAPTER 2

An Overview on Rumour Detection

In this chapter we would like to provide the reader a brief overview on the rumour detection research landscape. As aforementioned in chapter 1, most work focuses on identifying rumours at an aggregated level, referred to as macro-level. This means that a group of posts, clustered according to conversations, topic, enquiry content or event, is classified as rumour or not. However, these analyses do not distinguish whether the single posts in a group are actual rumours or not, an information that could be valuable in some domains, such as health or security. For this reason some researchers tackle this problem at the micro-level, i.e. considering the single posts classification. In the followings we are going to explore the literature specifying for each work the level of analysis and finally highlighting the motivations beyond this work.

2.1 Brief review of the literature

Computational analyses in the area of social media have started in the latest decades with a particular attention to social networks data and have provided a series of solutions usually named as rumour resolution and classification systems. According to the survey presented by Zubiaga et al. (2018), this framework comprehends four main modules: 1) rumour detection, that identifies which information constitutes a rumour or not; 2) rumour tracking, which focuses on monitoring the diffusion of the rumour recognising particular patterns; 3) rumour stance classification, that aims at identifying the orientation of posts discussing the veracity of the rumour; 4) rumour veracity classification, which is the final step of determining if the rumour news is actually true or false. In this work we focus on the first module, i.e. the rumour detection; next subsection therefore presents a summary of the background in this area, considering both traditional machine learning approaches and deep learning.

2.1.1 Rumour Detection in social microblogs

As we already mentioned, internet and social networks have granted researchers access to a massive quantity of data and real-time information, allowing the use of compu-

Rosa Sicilia

tational methods based on machine learning techniques to improve automatic rumour identification.

In this scenario, we can distinguish approaches built with traditional machine learning (Castillo et al. 2011, Agichtein et al. 2008, Alonso et al. 2010, Hughes and Palen 2009, Wu et al. 2015, Ma et al. 2015, Fard et al. 2019, Kwon et al. 2013, 2017, Zubiaga et al. 2017, Zhao et al. 2015) or based on deep learning techniques (Ma et al. 2016, Wang, Guo, Wang, Li and Tang 2019), which are now summarized in the next two subsections.

Traditional Machine Learning approaches

The main peculiarity of the traditional approaches is the use of hand-crafted features designed for the specific task. Castillo et al. (2011) faced a binary classification task distinguishing between rumour and non-rumour at the macro-level of Twitter topics (i.e. collections of posts and reposts). In particular the representation of the samples was inspired by previous work (Agichtein et al. 2008, Alonso et al. 2010, Hughes and Palen 2009), where each set of tweets was described with a selection of message-based, user-based, topic-based and propagation-based features, also adopted in later studies (Wu et al. 2015, Ma et al. 2015, Yang et al. 2012). The method was tested on a Twitter dataset collected between April and September 2010, considering only the newsworthy topics at the time. Their approach exploited a J48 decision tree with a three-fold cross validation, obtaining precision and recall in the range $70\%-80\%$. A different approach is presented by Wu et al. (2015), where the rumour detection is analyzed at the macro-level of conversations considering the Chinese social microblog Sina Weibo. In this case each set of posts belonged to the same conversational thread and it was described with new features related to the user, the message and the repost, in addition to those proposed in (Agichtein et al. 2008, Qazvinian et al. 2011, Yang et al. 2012)[1]. Then, the classification task was carried out using a hybrid Support Vector Machine (SVM), which integrated a traditional radial basis function with a random-walk graph kernel to model the propagation patterns of the messages in a tree. Tests were run on a dataset retrieved from the Sina Weibo community management centre, including about 5000 messages, 4 million users and many different topics. A three-fold cross validation attained precision and recall larger than 90%.

A different macro-level approach is presented by Zhao et al. (2015), where the authors presented a real-time method to identify rumour clusters of posts whose topic is a disputed factual claim. They started from the assumption that a rumour post would

[1]It is worth noting that Qazvinian et al. (2011) and Yang et al. (2012) did not deal with rumour detection on social media posts, but they presented a system tailored for rumour stance and veracity classification. The former dealt with data listed in About.com Urban Legends, while the latter used the Sina Weibo official rumour busting service.

Rosa Sicilia

generate questions and enquiries from the other users of the microblog. For this reason they manually created a list of regular expression to identify enquiry posts, such as "is (that|this|it) true", and then they clustered them according to content overlaps. Finally, they extract 13 statistical features from the clusters, such as the percentage of enquiry tweets, the average number of retweets, URLs and user mentions and the tweet length, and perform rumour classification using two learning paradigm, i.e. a Support Vector Machine and a Decision Tree. This method was tested on two Twitter datasets, one about the Boston marathon bombing and another on general topics retrieved in streaming, that were grouped in a total of 639 clusters for rumour detection. On these data they were able to achieve above the 70% of precision in rumour classification.

For what concerns the micro-level analysis, Zubiaga et al. (2017) presented a system able to classify each post as rumour or non-rumour while relying on its content in conjunction with context learnt from earlier posts associated with the same event. The authors extracted content-based and social features from each post and modeled a given timeline as a linear chain. This was given as input to Conditional Random Fields to classify the current post based on the sequence of rumours and non-rumours preceding it. The method was tested on a two-class Twitter dataset comprising five different newsworthy events and filtering out those tweets that had less than 100 retweets because they are interested in applications where tweets provoke a high number of retweets. The final dataset accounted a total of 5802 samples: 2079 for "Charlie Hebdo" event, 1143 for "Ferguson" event, 469 for "Germanwings Crash" event, 890 for "Ottawa Shooting" event and 1221 for "Sydney Siege" event. The samples were labeled in two classes, namely rumour and non-rumour, and the final tests led to above 90% of recall on rumours. In a more recent work Fard et al. (2019) address the micro-level rumour detection problem considering a one-class approach. In particular, the authors extracted 86 features related to the linguistic and to the content, to the user and to the meta-data linked to each post. Then they compared the performance of seven different one-class classification algorithms trained only with the rumour posts of two Twitter datasets, for a total of more than 150000 tweets. The final performance was measured in terms of F1-score, that ranged from 60% to 94%.

Beyond these contributions that exploit representations focused on the analysis of user characteristics and message propagation patterns, other work has addressed the problem of catching the temporal dimension in rumour diffusion (Kwon et al. 2013, Ma et al. 2015, Kwon et al. 2017). In detail, Kwon et al. (2013) turned the attention on the temporal aspects of rumour diffusion at the macro-level of topics analysing Twitter social microblog. The developed model was based on periodic time series that showed that rumour life-cycle is characterized by fluctuations such as daily and external shock cycles. This temporal model together with the structural and linguistic aspects of the

Rosa Sicilia

posts was used to extract a pool of features used to compare different classification algorithms. Tests on a private Twitter dataset, accounting 130 topics, yield to the best accuracy of 90%. Differently, Ma et al. (2015) tried to catch the temporal variation of social context features due to the message spread over time. The posts were aggregated according to specific events (macro-level) and each post was associated with a timestamp, computed considering the moment the microblog was posted and the delay between the first and the last post published about the same event. Then, for each of these time stamps, the authors computed the variation of the features, using the slopes between two consecutive intervals. Test phase was conducted on the general purpose dataset presented by Castillo et al. (2011) and Wu et al. (2015), yielding an accuracy that ranges from 80% up to 90% for the early stage rumour detection. A more recent analysis presented by Kwon et al. (2017) explored the differences between rumours and non-rumours over varying time windows, to pursue early stage as well as long term detection. The authors examined a comprehensive set of user, structural, linguistic and temporal features extracted from a public Twitter dataset accounting 111 events labelled at the macro-level. Finally, a Random Forest classifier was tested in 3-fold cross validation over different time windows, attaining accuracy ranging from 60% up to 90%.

Deep Learning approaches

In contrast with the traditional machine learning approach that exploits hand-crafted features, researchers have recently started to explore the potential of deep learning for rumour detection. Ma et al. (2016) developed a model based on Recurrent Neural Networks (RNN), aiming at learning hidden representations for the variation of contextual information. The rumour detection task was handled considering directly the veracity prediction at the macro-level of events. Each event was divided into time intervals to create variable length time series that were fed into different RNN models. The model was evaluated on a dataset of 498 rumour and 494 non-rumour events from Twitter, and 2313 rumour and 2351 non-rumour events from Sina Weibo, obtaining the best accuracy equal to 91%. Besides the temporal variation, Wang, Guo, Wang, Li and Tang (2019) presented a novel two-layer gated-unit RNN (GRU) which is combined with a dynamic time series model and a sentiment dictionary that catches the opinion-related properties of the data. First, each post is aggregated at the macro-level of events, and each event is divided according to a fuzzy clustering algorithm into time intervals. Second, the sentiment dictionary is applied to this representation of data and finally given as input to the GRU network. The model was evaluated on a public dataset of 2313 rumour and 2351 non-rumour events collected from Sina Weibo, yielding to the best accuracy equal to 96%.

Rosa Sicilia

2.2 Staring point of our Rumour Detection research

The overview of the literature presented so far shows that the large part of the work has dealt with the macro-level analysis investigating if a topic, an event, a cluster or a conversation contain unverified news, without studying if a specific post within those sets is a rumour or not. As mentioned in section 1.2, in many applications, such as health or security, the possibility to know exactly which posts are rumours within an event or a conversation could be of paramount importance. On the one hand, it could be necessary to immediately know the exact source of a rumour, on the other hand, it is often possible to find rumour and non-rumour information belonging to the same conversation or topic. Nevertheless, to the best of our knowledge, two pioneering work have started considering the possibility to address the issue at micro-level (Zubiaga et al. 2017, Fard et al. 2019). However, each suffers from the following limitations. First, the approach designed in Zubiaga et al. (2017) is limited to recognize rumours posts that have a high diffusion rate, i.e. they have to be retweeted at least 100 times. Although this choice is consistent with the fact that rumours can attract substantial interest among the users, it is a limitation of the method since it could miss newly rising rumours that haven't reached the maximum spread, yet.

Second, Fard et al. (2019) considers a one-class classification problem in order to detect only rumour posts in the network, arguing that definitions of a "non-rumour" class are ambiguous and could be misleading in the rumour detection task. On the contrary, we deem that interpreting "non-rumours" as a referenced news and being able to distinguish them could have value in applications such as health where it could be important to know which information is the reliable and which is not. Indeed, as it is extremely immediate, people are prone to rely on knowledge and advices found online and mainly on social microblogs, where it is effortless to search and share everything right away. This happens for any possible news, but it could have considerable impact in people lives when it is related to health knowledge. It is worth stressing that not all the resources provide accurate or reliable information (Wang, McKee, Torbica and Stuckler 2019), making health news a straightforward application for a micro-level rumour detection system.

For the aforementioned reasons, we decided to investigate rumour detection at the micro-level and within the specific health topic domain. We considered also posts with a low diffusion rate in order to be potentially able to detect newly rising rumours and a multi-class problem to have a fine grade of detection. We deeply analysed the feature representation of the posts, the reliability of such descriptors applied to different domains, how to find the best representation and also the knowledge transfer among health-related topics.

Rosa Sicilia

CHAPTER 3

Twitter Health-related Datasets

Twitter is a social microblog that counts millions of users from all over the world, facilitating real-time propagation of information to a large group of people and allowing users to carry out different actions, such as posting 280-character-long messages (tweets), replying to such posts (replies), or forwarding them (reposts or retweets). These simple actions facilitate real-time propagation of information to a large group of people, making it an ideal environment for rumours diffusion. A set of tweets belonging to the same topic can be grouped by keywords, such as hashtags. To cope with the challenge of rumour detection in the health domain we hereby present two Twitter health-related dataset.

3.1 The Zika virus dataset

This work copes with rumour detection in health-related news, a topic that to the best of our knowledge, has not been specifically investigated yet. For this reason a public dataset was not available and we therefore queried Twitter using the keyword *#zikavirus*, which was one of the main sanitary trends in 2016. Indeed, Zika virus disease was declared a Public Health Emergency of International Concern (PHEIC)[1] by the World Health Organization on February 1^{st}, 2016. We retrieved 497 posts in two days of April 2016, and 1085 posts in three days of May 2016 with a gap between the two time intervals. Even if such time intervals could seem short, they ensure that the spreading of the news throughout the network is not the mere outcome of homophily, but the result of a real interest of the users in that specific news (Dave et al. 2011). Moreover, the gap between the first and the second acquisition ensures that the dataset contains different rumour tweets: indeed, if we considered only tweets collected in consecutive days, the dataset could contain many repetitions of the same post as rumours show a bursting behaviour tending to appear and to fade repeatedly in a few days (Kwon and Cha 2014, Kwon et al. 2013). The infectious nature of the rumour suggested us to consider not informative, and therefore to discard, the tweets that did not generate any retweet or reply and the tweets whose propagation graphs could not be reconstructed.

[1] About Zika Virus Disease: http://www.cdc.gov/zika/about/

Rosa Sicilia

Finally, the dataset contains 709 samples that were manually labelled by two human annotators into three classes, namely *rumour* ($\sim 54\%$ of samples), *non-rumour* ($\sim 30\%$ of samples) and *unknown* ($\sim 16\%$ of samples). Each tweet was blindly annotated and possible disagreements were solved through debate for a common understanding. The annotators were also made aware of the scope of the work, so they knew the following definitions of the three classes (DiFonzo and Bordia 2007):

- ***rumour***: as presented in section 1.1, a rumour is an unverified and instrumentally relevant information statement in circulation. It indicates news without reference, so that can't be verified, and that is around. Typically it shows an instantaneous and huge spread in the network, in other words an *infectious behaviour*. An example retrieved from the dataset is:

  > *RT @ClassicPict: Mosquitoes kill more annually than Sharks. #ZikaVirus https://t.co/P2xi13Vx2Z.*

- ***non-rumour***: this class identifies all referenced news with at least one link to a certified and official webpage, such as news papers, hospitals, universities, etc.. Below there is an example of a tweet linked to the WHO webpage.

  > *RT @WHO is dispelling rumours around Zika* and *microcephaly. See the facts about #ZikaVirus: http://goo.gl/JDKuys*.

- ***unknown***: this class collects all the news which cannot be determined to be false, such as news with a link to an empty page, news potentially true but without reference, news with a main topic unrelated to the one of interest. An example of potentially true news without any reference is:

  > *RT @DoughertyNews9: ICYMI: Researchers discover #ZikaVirus can be carried by a 2nd species of mosquitoes. We have details soon on @NEWS9*

3.2 The Vaccine dataset

For this dataset we collected a total of 1870 tweets in June 2018, using the health-related keyword *#Vaccine*. After that, the posts were labelled following the classes definition presented in the previous section, where we identified three categories, namely *rumour*, *non-rumour*, and *unknown*. In this case the labelling process was blindly carried out by three human annotators, and their labels were later used to build a gold standard computed by the majority voting on the complete set of labels of annotators involved in this study. In those cases where there was total disagreement between annotators (i.e. three different labels for the same post), the samples were discarded (a total of 107 samples). Moreover, since in the retrieved data there were many empty posts or tweets reporting

Rosa Sicilia

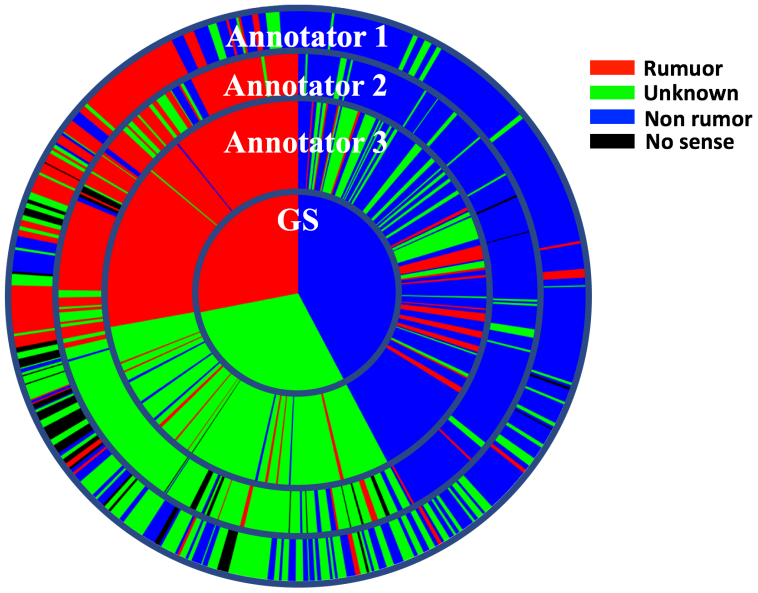


Fig. 3.1 Pie chart representing the distributions of labels assigned by each annotator to each sample, which are compared with respect to the Gold Standard (GS).

the research hashtag but actually about a topic different from the one of interest, we decided to include a fourth class, named *no sense*, to collect this noisy data. Samples that had a gold standard label of *no sense* were also left out the following analysis (a total of 354). The final dataset accounted 1409 samples, with 28% of rumours, 30% of unknowns and 42% of non-rumours. Figure 3.1 offers a graphical representation of the distribution of labelled samples among the three annotators with respect to a gold standard. Round angle of the circle is divided into 1763 parts (1409 final samples + 354 no sense), and each circle sector corresponds to a sample. The figure visually shows a certain degree of variability between the labels provided by each annotator.

Rosa Sicilia

CHAPTER 4

The Novel feature set

As first step to tackle the micro-level rumour detection in single topic domain, we focused on the data representation, trying to identify descriptive measures able to characterize the data. We hereby present a machine learning-based rumour detection system that leverages on newly designed features, including influence potential and network characteristics measures. We tested our approach on two real-world datasets composed of health-related posts collected from Twitter microblog. The rest of the chapter is organized starting from the features design (sec. 4.1), then we study the descriptors evaluation and selection (sec. 4.2), and finally we present a further analysis on their robustness validation (sec. 4.3).

4.1 Feature Set Design

We hereby present a machine learning-based rumour detection system whose pipeline is shown in Fig. 4.1. The upper panel depicts the training phase: first the system acquires the data from Twitter, then it identifies two structural levels (named as *user* level and *network* level in the following) that lead to the extraction of representative features. Feature evaluation and feature selection on a disjoint validation set permit to build the learning model, which is finally trained. To perform the classification on a new set of tweets, the lower panel of Fig. 4.1 shows that the system exploits both the best feature set and the learning model defined in the training stage.

Differently from deep learning approaches, in traditional machine learning a fundamental step is the data representation hand-crafted design. In this section we will focus on the core of the chapter which is the feature set design that will be evaluated and validated following the aforementioned pipeline.

4.1.1 Twitter Structure

Fig. 4.2 shows a representation of the Twitter structure, which can be divided into user and network levels.

The former identifies the characteristics of the user and his/her statuses, which are

Rosa Sicilia

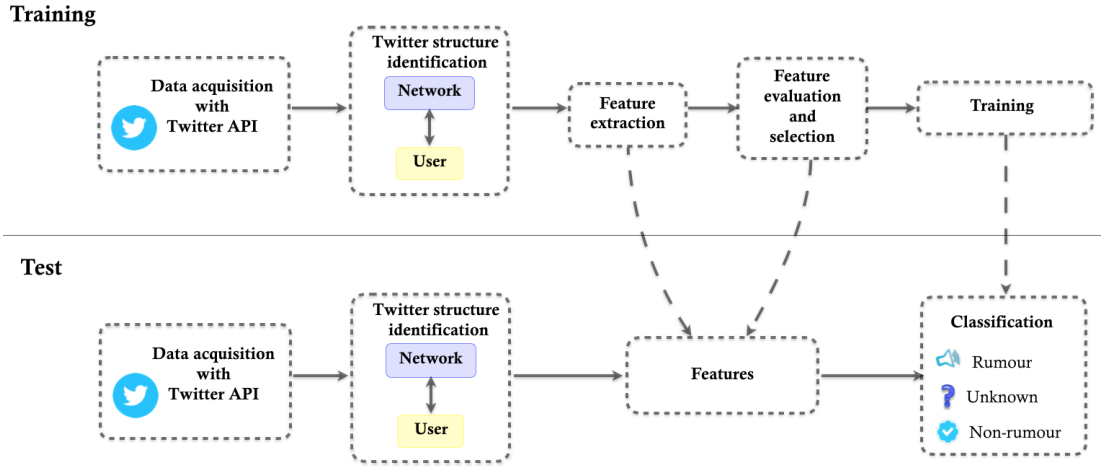


Fig. 4.1 Pipeline of the rumour detection process.

considered as self-standing and without links in the network. For instance, some properties observable at this level are the number of followers and followings of a single user or the sentiment related to a specific post.

The latter identifies the interaction between users in the network and the related properties given by retweets and replies. Such two actions can be represented by conversation graphs, where each node denotes a user and each link between the nodes corresponds either to a retweet or to a reply. Fig. 4.2 shows two examples: on the left there is a retweet graph with a star shape, where each action connects a retweet (red node) to its original post (grey star). On the right there is a graph of replies, which has a more intricate structure with an original post leading to multiple intermediate points (blue nodes) and multiple end points as well (green nodes). The characteristics of the conversation graphs allow to analyse the spread of information in the network and the influence properties of the users.

4.1.2 The features set

We now present the set of features, which has been divided into three different groups: referred to as *influence potential*, *personal interest* and *network characteristics* (Table 4.1). Each group catches a different property of the social microblog, and within each group measures related to both users and to the network levels can be clearly distinguished. The interested readers can refer to the appendix A for the detailed feature definition.

Influence Potential Measures. These measures describe the power or the capacity of causing an effect in indirect or intangible ways. Some user and network level features were introduced by Castillo et al. (2011) and by Wu et al. (2015), where the authors suggested to characterize a user considering all the properties related to his/her Twitter account. Therefore, we computed the number of followers and followings of the user,

Rosa Sicilia

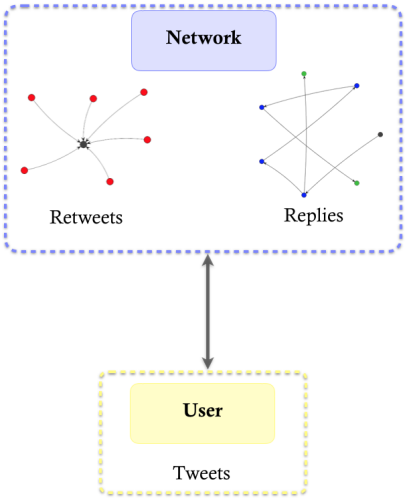


Fig. 4.2 Representation of the Twitter structure. The retweet conversation is represented by a star graph with all retweets (red nodes) referring to an original post (grey node). Conversely, the replies show a more complex structure, with an original post that leads to multiple intermediate points (blue nodes) and to multiple end points as well (green nodes).

as well as if he/she was follower of another user involved in a conversation, and the age of his/her account, defined as the difference in months between the registration date and the current date. We considered also features related to the specific tweet (presence of URLs or question marks, etc.). Nevertheless, we also introduced two new user level features named as P_{rt} and P_{url}. P_{rt} is the probability that a tweet is retweeted, computed as the number of times the tweet is shared (if it is a retweet) divided by the total number of samples in the dataset. P_{url} is the probability of a URL to be shared; it counts how many times the attached URL was used as a reference source divided by the total number of URLs in the dataset.

Personal Interest Measures. Understanding the reaction of people to a specified news in terms of opinions and related sentiments is a fundamental step for rumour detection: in fact they convey much information about what people actually think of a specific news, if they believe it or not. To extract this information from the retrieved tweets we performed a lexicon-based sentiment analysis extracting the SentiWordNet score (Esuli and Sebastiani 2006), including a stage of negation detection to take into account the presence of negated contexts. For example, if a status says *"I don't like this post"*, the score of the word *like* switches from a positive to a negative value, because of the presence of the negation *don't*. The sentiment score is computed summing up the scores for every match of each token in the tweet with a word belonging to lexicon synset, considering that all the negation words are changed in sign. While the sentiment score is directly used as user level feature, network level features were computed considering the scores assigned to all tweets belonging to each conversation (Castillo et al. 2011).

Rosa Sicilia

Feature group	User Level	Network Level
Influence Potential	#Followers #Followings #Statuses Registration Age is a Retweet? (isRT) has URL? (hasURL) has "?"? P_{rt}† P_{url}† is Follower?	Avg #Followers Avg # Followings Avg # Statuses Avg Registration Age
Personal Interest	Sentiment score	Avg Sentiment score Fraction of positive Fraction of negative
Network Characteristics		PageRank◇ Closeness centrality◇ Betweenness centrality◇ Conversation size† Fraction of Users followers of the root (FUFR)† Fraction of tweets with URL (FTWU)†

Table 4.1 List of all the features extracted, where "# " and "Avg" stand for "number of" and "average", respectively. Newly introduced features are marked with a dagger (†). The features inspired by graph theory that, to the best of our knowledge, have not been used in rumour detection yet are marked with a diamond (◇).

Network Characteristics Measures. These features were designed to describe the characteristics of the propagation graphs built with retweets and replies conversations. To identify how the information spreads in the network we quantified the degrees of influence of the users. To this aim, we considered measures of centrality and popularity originally conceived for the graph theory, such as the PageRank, an index of the popularity of the node in a network, the closeness centrality, a measure of the independence or efficiency of a node, and the betweenness centrality, a representation of the potential of a point for control of communication (Freeman 1978). We also introduced other three structural measures: (i) the conversation size, which is the number of nodes in a conversation, (ii) the fraction of followers of the root, given by the number of users in a conversation who are followers of the root user, divided by the conversation size, (iii) the fraction of tweets with an URL, computed as the number of tweets with a URL in a conversation divided by the conversation size.

4.2 Feature Evaluation and Selection

The evaluation of the designed measures is an important step in the process of feature engineering: it aims at analysing which features are the most informative and whether the newly introduced ones are meaningful or not for the classification purpose.

Among the different techniques for feature evaluation (Huang 2015), we applied the wrapper method that compares different candidate sets of variables using the performances of classification algorithms. We analysed every feature by evaluating how

Rosa Sicilia

the Area Under the ROC curve (AUC) changes when it is removed from the feature set. We used the AUC since it is independent to the decision threshold, and it is invariant to the a priori class probability distribution. The discriminant power of a feature is measured trough the following score:

$$S(f) = AUC_{complete\ set} - AUC_{leave\ f\ out\ set} \tag{4.1}$$

where f is the considered descriptor in the feature set. Hence, the larger the value of $S(f)$, the more discriminant the feature f.

To avoid any bias given by the specific classification algorithm used, we ran such an evaluation using several classifiers belonging to different learning paradigms (listed in section 4.2.1). Feature evaluation was conducted in 10-fold cross validation on the training set and the results were then summarized via a rank analysis computing the relative performances of one feature with respect to the others. First, for each classifier used, we sorted individually the values of $S(f)$ and then we assigned to each variable f a rank with respect to its place among the others. The highest rank is 24 (assigned to the worst case corresponding to the least discriminant feature) and the lowest is 1 (assigned to the best case). Ranks of each feature are finally summed up per classifier, and then they are normalized with respect to the highest possible value (i.e. highest rank $\times$ numbers of classifiers).

The feature selection step aims at finding the most representative measures among all those computed. To this end, we first applied a threshold on the aforementioned ranks returning a set of features. Second, on such a reduced feature space, a classifier is trained on the training set. Third, this classifier labels samples belonging to an independent validation set, where we measured the AUC. Such a three-steps procedure is repeated for different values of the threshold so that, at the last iteration, the feature set is composed only by one descriptor. The selection is then repeated for different learning paradigms and we find out the reduced feature set providing the largest average AUC per classifier. Next, we discuss the experimental results.

4.2.1 Experimental Setup

This first analysis for feature evaluation was conducted on the Zika virus dataset, which was divided into training (90% of samples) and validation (10% of samples). Assuming that classification tests are described as Bernoulli-processes this partition ensures that, when estimating the performance, the width of the 95% confidence interval is equal to 0.1, which is a reasonable choice.

We considered classifiers belonging to different learning paradigms, including a Multi-Layer Perceptron (MLP) as a neural network, a Nearest Neighbour (NN) as an instance-based classifier, a Support Vector Machine (SVM) as a kernel machine, a Ran-

Rosa Sicilia

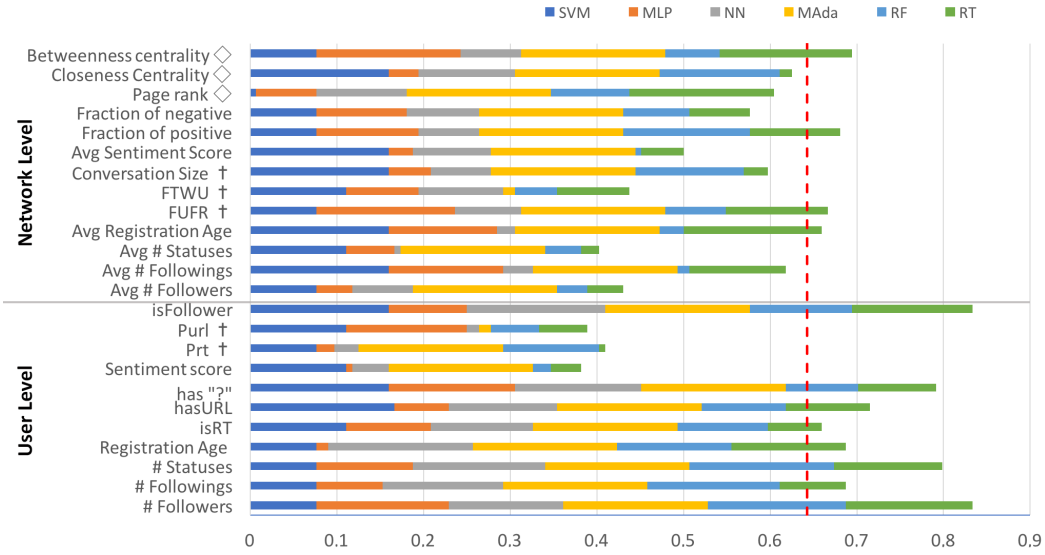


Fig. 4.3 Stacked histogram of the rank analysis, where the shorter the bar the more informative the feature is. The dotted red line is the median rank value. The bars also show the contribution of each classification algorithm to the importance of the single variable, highlighted with different colours. The newly introduced features are marked with a dagger (†), while features inspired by graph theory are marked with a diamond (◇).

dom Tree (RT) as a decision tree, a multiclass Adaboost (MAda) as a multi-expert system and a Random Forest (RF) as an ensemble of trees. We set their parameters to the default values. Although we acknowledge that their tuning could lead to better results, we preferred to maintain a baseline configuration as the basis for comparison between them. Indeed, it is reasonable that the classifier which wins on average on all the experiments would also win if a better setting was performed (Fernández et al. 2013). Furthermore in a framework where the classifiers are not tuned, the winning learning model tends to correspond to the most robust one, which is also a desirable characteristic.

Next subsections describe the experimental results of the feature evaluation stage and of the feature selection phase, and finally subsection 4.2.3 reports the performances achieved in the testing scenario.

4.2.2 Evaluation and Selection results

Fig.4.3 shows the ranks of each feature f computed as described in section 4.2. Each bar in the plot is the normalized sum of the ranks provided by different learning paradigms, whose contributions are represented in different colours. Note that the shorter the bar the more informative the feature is.

The newly introduced features (marked with a dagger) show considerable descriptive power compared to the others as their ranks are usually lower than the median.

Rosa Sicilia

In particular the user level measures P_{rt} and P_{url} as well as the network level feature FTWU are worthy of mention. The P_{rt} value conveys information about how much "popular" a node is, computing its probability of being retweeted. Since, usually, the rumours present a huge spread in the network, these tweets have a high value of the P_{rt}, thus helping the distinction between rumours and non-rumours. P_{url} represents the probability of a URL to be shared in a post: its informative value lies in the strong habit of users to support a statement linking a web page to the post. Therefore, if this score is high, it means that the considered tweet is probably a non-rumour, and vice versa. The FTWU, measuring the fraction of tweets with a URL attached, appears particularly informative for the problem. Indeed, people manifest their disagreement with a not reliable news using some different referenced sources.

Among the newly introduced features, the closeness and betweenness centrality measures attain high rank values (Fig. 4.3), suggesting that they are not so informative. By definition, such features discriminate less when computed on very small networks (e.g. a two node network), which frequently occur in our dataset. This happens since we adopted a general approach where small conversations and not influential users were not filtered out to favor the micro-level analysis.

We also note that the sentiment score attains a low rank. This finding agrees with data reported in the literature (Castillo et al. 2011, Wu et al. 2015, Yang et al. 2012, Qazvinian et al. 2011), showing that the evaluation of both the sentiment in single tweets and the average sentiment in a conversation is particularly important for the rumour detection problem. That's because the information related to the opinions can be highly discriminative, especially if we consider the distinction between neutral posts and tweets with high scores (the absolute value of both negative and positive sentiments). In fact, it's more likely that a neutral status simply reports news with its reference, whereas emphatic statuses try to gain the attention of users in the social network, and use the infection to spread unreliable news.

The analysis of the results also shows that the most informative features are actually those belonging to the network level since almost all of them have a normalised rank lower than the median. Indeed, except for P_{rt}, P_{url} and for the sentiment score, we found that the user level attributes (which related to the single nodes of the network) do not discriminate well the samples in the feature space. Although this result disagrees with previous findings reported in the literature, we deem this happens because we do not delete the small conversation, as most of the previous work do (Castillo et al. 2011, Wu et al. 2015, Ma et al. 2015). In fact, such work considered only those networks where the news spread was strongly dependent on the nature of the nodes (e.g. with a high level of opinion leadership (Wu et al. 2015)). In this case the features related to the nodes represented valuable information for the rumour detection task. Consequently

Rosa Sicilia

the networks originating from a node with a low level of opinion leadership were all filtered out. Unlike what reported in the literature, the presence of small conversations in our dataset allows us to maintain the classification problem as much general as possible, so that the classifier cannot discriminate the spread of a rumour only from the characteristics of the single agents of the networks.

The feature selection experiments were conducted as described in section 4.2 on the basis of the results showed in Fig.4.3.

We estimated the performances of each classifier on the validation set and then we averaged the results among all the classifiers. The results show that the best feature subset is composed of 20 features, discarding the *number of followers*, the *number of statuses*, *has "?"?* and *isFollower*, which also showed the highest ranks (Fig.4.3). This finding agrees with the observations presented earlier in this section: the measures related to the user's characteristics have been discarded, in that they provide little information.

4.2.3 Classifier Performances

As preliminary analysis for the overall rumour detection system presented in Fig. 4.1, we hereby look for the best learning paradigm for our classification problem also considering the system validation on an independent dataset.

To determine the best classifier we compared the learning paradigms mentioned in section 4.2.1 running a 20-fold cross validation on the training set, representing the samples in the feature space selected as described in section 4.2. As performances indexes we measured the overall accuracy and the average AUC per class[1]. Table 4.4 shows the results of such an exhaustive comparison: in each panel the lower triangle of the tabular shows the amount of win-tie-loss of a classifier in a row comparing with a classifier in a column, whereas the upper triangle reports the p-value of the pairwise difference between the classifier performances measured using the Wilcoxon rank sum test. In panel a it is straightforward to observe that the RF significantly outperforms most of the other learning paradigms: indeed, the p-value is many times smaller than 0.01 and the number of wins is close to 20, i.e. the maximum value. Similar observations hold also in case of AUC (panel b of the table), where in particular RF wins over all the other classifiers. Note also that RF, as well as SVM, has been the most used in the literature on rumour detection (Castillo et al. 2011, Yang et al. 2012, Wu et al. 2015, Ma et al. 2015).

We also validated our rumour detection system computing the performances on a

[1] A performances metric is named "per class" when it is computed considering the binary classification task derived from the decomposition of the original recognition task.

Rosa Sicilia

	SVM	MLP	NN	MAda	RT	RF	Avg acc (%)
SVM	-	9.8E-03	1.8E-04	8.2E-01	8.4E-06	4.8E-06	71.96
MLP	15-3-2	-	1.6E-01	1.8E-02	1.8E-02	3.4E-03	78.59
NN	17-0-3	13-3-4	-	4.5E-04	2.9E-01	6.2E-02	81.58
MAda	3-16-1	2-4-14	3-1-16	-	2.3E-05	9.2E-06	72.42
RT	19-1-0	15-3-2	12-3-5	19-1-0	-	3.2E-01	84.09
RF	19-0-1	18-1-1	14-4-2	19-0-1	10-7-3	-	86.13

Table 4.2 Win-tie-loss and p-value comparison in terms of accuracy (acc).

	SVM	MLP	NN	MAda	RT	RF	Avg AUC (%)
SVM	-	3.4E-07	6.6E-06	5.2E-01	2.3E-06	6.7E-08	73.63
MLP	20-0-0	-	7.9E-01	1.2E-06	9.1E-02	2.5E-04	89.06
NN	18-0-2	10-0-10	-	2.3E-05	1.7E-01	8.3E-05	87.86
MAda	14-0-6	0-0-20	3-0-17	-	1.2E-05	6.8E-08	74.92
RT	19-0-1	6-0-14	7-0-13	19-0-1	-	3.5E-06	85.66
RF	20-0-0	20-0-0	19-0-1	20-0-0	20-0-0	-	95.65

Table 4.3 Win-tie-loss and p-value comparison in terms of AUC.

Table 4.4 Exhaustive comparison between the performances of different classifiers expressed in terms of accuracy (panel a) and AUC (panel b). In the lower triangle each tabular shows the amount of win-tie-loss of a classifier in a row comparing with a classifier in a column. The upper triangle reports the p-value of the pairwise difference between the classifier performances measured using the Wilcoxon rank sum test. Last column of the table reports the average values of accuracy and AUC for each classifier computed running a 20-fold cross validation on the training set.

test set using the best feature subset and the RF classifier, as suggested by the aforementioned results. The test set was collected using the same procedure described in section 3.1, and it consists of 91 labelled samples, where the a-priori probabilities of rumour, non-rumour and unknown are equal to 55%, 31% and 14%, respectively. It is worth observing that such a set contains a number of samples almost equal to the validation set introduced in section 3.1, thus ensuring that the confidence interval of estimated performance is the same.

We obtained the following performances: the overall accuracy is equal to 73.63%, the average precision per class is 72.80%, the average recall per class is 73.60% and the average AUC per class is 89.00%. Furthermore, as we are mainly interested in the recognition of rumour samples, which we regarded as "positive" samples, the true positive rate and the false positive rate achieved on the rumour class are equal to 92.00% and 39.00%, respectively. This means that the system presented so far is *liberal*, i.e., it tends to classify samples as rumour, giving sometimes false alarms that could be tolerated for the purposes of this work. Indeed, in the health tweet domain it is reasonable that the cost of a false positive is lower than the one of false negative. This suggests that

Rosa Sicilia

the use a liberal system should be advocated in all those delicate cases where the spread of an unreliable information is far more dangerous than questioning the trustworthiness of official statements.

Finally, it is worth noting that the performances reported in this work cannot be compared with those presented in the literature since the recognition task defined here is different from those previously defined by Zubiaga et al. (2017) and Fard et al. (2019). Indeed, on the one hand Zubiaga et al. (2017) designed a rumour detection system for sequence of posts that have at least 100 retweets, so excluding those with a low diffusion rate. On the contrary, these are included in our dataset, since we plan to recognize also rumours that have not spread, yet. Hence we deem that their method is not directly comparable to our system. On the other hand, the work by Fard et al. (2019) models the rumour detection task as a one-class classification problem, which is also not comparable with our three-class approach.

4.3 Feature Robustness Validation

The analysis presented so far has shown the representative power of the feature set on one health-related dataset, but we deem that a necessary validation step is studying the robustness of the feature set when it is applied to data coming from different topics. In fact, there is the risk of having features that exhibit a domain-specific distribution (Tolosi et al. 2016). This clearly could introduce a bias in the developed system that hinders its generalization capability to other domains.

4.3.1 The issue of overfitting

It is well known that the issue of generalization is an hot topic in machine learning and pattern recognition, which has attracted much research (Bishop 2006, Vidyasagar 2002). Many researchers have investigated this issue from the perspective of overfitting the training data. Learning approaches can face also with an even more difficult generalization challenge: how much an algorithm trained on some data is able to recognize samples belonging to a similar, but not the same, domain? In fact, in the case of Twitter rumour analysis, biases introduced by features that exhibit a domain-specific distribution could hinder the prediction capability of the overall system on unknown samples belonging to another domain. This can straightforwardly drop the performance when compared to that achieved in a controlled test environment (i.e. when the tests are performed in cross-validation on a given dataset).

Although the issues of overfitting and feature robustness related to generalization can also be approached from the perspective of semi-supervised learning, in the following we will analyse such topics in a supervised fashion. In particular in the following

Rosa Sicilia

we will test the pipeline presented in Fig. 4.1 on the Vaccine dataset described in section 3.2.

4.3.2 Experimental setup and results

In order to analyse the feature robusteness to overfitting we extecuted two experiments:

EXP#1 In the first experiment we tested the pipeline reported in figure 4.1 with a 10 fold stratified cross validation on the Vaccine dataset, using the gold standard as ground truth.

EXP#2 The second experiment evaluated the pipeline in figure 4.1 with a 10 fold stratified cross validation on the Zika virus dataset.

The feature evaluation and selection was performed using ReliefF, a simultaneous feature ranking method that calculates relevance weights for all the features at the same time by looking into their joint relationship with the target (Huang 2015). Then, we used the Random Forest ensemble as learning paradigm, since it is one of the best performing for rumour detection problems (Zeng et al. 2016). It is worth noting that this approach differs from the one presented earlier (sec. 4.2): indeed, in the previous section we used a wrapper evaluator to find the best feature subset and to analyse the importance of each descriptor using several learning paradigms. On the other hand, we hereby turn the attention to the representative power of features over different datasets, so we used a ranking feature evaluator independent from the classification algorithm.

Results for EXP#1 and EXP#2 are presented in Table 4.5, where our features set shows outstanding performance on the #Vaccine dataset, with a percentage of correctly classified rumours equal to 95.2% versus the 88.4% of #Zikavirus. On the one hand this disparity could be related to different structural characteristics of the two domains, having that #Zikavirus rumour news are actually more difficult to analyse and detect than those related to vaccines. On the other hand, this issue could be also linked to the difference in the annotation process for the two datasets: indeed, the #Zikavirus data was blindly labelled by two annotators and possible disagreements were solved through debate for a common understanding, so without keeping trace of the individual annotations, whereas for the #Vaccine data we were able to build the gold standard described in chapter 3. However, due to the complexity and time-consuming effort of a manual annotation process, we had to leave the study of a gold standard for #Zikavirus data to future work.

Besides this, it is worth pointing out that results of EXP#2 are also different from those achieved in section 4.2.2 on the #Zikavirus data: in fact, hereby we obtained an overall accuracy of 82.3% versus the 73.6%. Despite this, the recall on rumour class in section 4.2.2 was 4% higher than in EXP#2. These variations are due to the

Rosa Sicilia

	Rumour class	**All classes**		
	recall	**acc**	**prec**	**F1**
EXP#1	.952	.960	.961	.958
EXP#2	.884	.823	.785	.767

Table 4.5 Random Forest performance, computed on #Vaccine dataset (EXP#1) and on #Zikavirus dataset (EXP#2). The first column reports the recall for rumour class. The "acc is a shorthand for accuracy, whilst "prec" stands for the average precision per class and "F1" is the average F1 score per class.

	Precision per class			**Recall per classes**			**Accuracy**
	Rumor	Unknown	Non rumor	Rumor	Unknown	Non rumor	
Annotator 1-GS	0.55	0.52	0.86	0.88	0.63	0.75	0.71
Annotator 2-GS	0.76	0.86	0.92	0.93	0.78	0.92	0.88
Annotator 3-GS	0.98	0.91	0.65	0.80	0.73	0.96	0.82
Average	0.76	0.76	0.81	0.87	0.71	0.88	0.81

Table 4.6 Annotators performance with respect to Gold Standard (GS), computed on #Vaccine dataset (EXP#1). The first tree columns reports the precision per class. The first tree columns report the recall per class is a shorthand for accuracy, whilst the fourth column reports the accuracy and the last the average values.

aforementioned difference in the analysis approach: in fact, in previous sections we focused on feature evaluation and selection, using a wrapper evaluator and a portion of samples as independent test, whereas here we wanted to study the feature robustness on different datasets considering a feature selection process independent from the learning paradigm.

Finally, we compared human average performance (computed over the three annotators scores in Table 4.6) and the system performance on the #Vaccine dataset. We observe for both overall accuracy and recall on rumours, that the system significantly outperforms an average human: in fact, results reported in table 4.5 show that the accuracy and rumour recall for #Vaccine data are 96.0% and 95.2%, respectively, versus 81.0% and 76% of the average human performance reported in Fig. 4.6.

4.4 Remarks

In this chapter we presented and evaluated a new set of features for rumour detection at the micro-level in a single topic domain related to health news using Twitter data. The restricted scale made the detection problem different from the literature, as it excludes the possibility of using the topic information as a feature, of making any a priori assumption on the nature of the network, and of making a priori assumptions on user's characteristics.

The feature set exploited not only features available in the literature, but also new

Rosa Sicilia

descriptors inspired by the study of the graph theory and the social influence models including the likelihood that a tweet is retweeted and the likelihood of a URL to be shared, conversation size, fraction of users followers of root, and fraction of tweets with URLs. Experimental study on the Zika virus dataset demonstrated the effectiveness of this system and new features which can correctly detect about 90% of rumours, with acceptable levels of precision. Moreover, we explored the selection of different features using different classification methods for rumour detection which is beneficial to the future studies in selecting effective combination of classifiers and features in such task.

As further validation, we presented an analysis of the robustness and generalization strength of such feature set, on a second Twitter dataset about vaccines. This was retrieved and manually labelled by three annotators, allowing us to determine a gold standard that was used for supervised learning of the model. We performed 2 experiments in order to investigate the robustness of the descriptors, analysing how such a feature set performs when trained and tested on a new topic of the health domain. Results show outstanding performance on the new data, indicating that the features hold a good discriminative power in detecting rumour for a specific topic data set. Indeed, the system significantly outperformed and average human annotator on the Vaccine dataset, reaching an overall accuracy of 96% (tab. 4.5).

Rosa Sicilia

CHAPTER 5

A New Filter for Feature Selection

In the previous chapter, we investigated rumour detection at the individual posts level and within a specific topic domain. As main contributions we introduced newly designed features, including measures based on influence potential and network characteristics, which were tested on two real-world datasets collecting health-related posts from Twitter microblog with the keywords #Zikavirus and #Vaccine. We also improved the system configuration for the #Zikavirus dataset, attaining an 82% of accuracy; furthermore, the application of the same features to a different topic (#Vaccine) yielded an accuracy equal to 96%, showing the robustness of the proposed descriptors to dataset variation.

To further investigate the impact of the feature selection stage in the proposed system, the research question that we advance in this chapter is *how does feature selection influence the system proposed so far?* To answer this question we introduce a novel method for feature subset evaluation based on a new algorithm that performs a rule-based feature space characterization. In addition we study the influence of feature selection in our rumour detection task comparing the proposed method with other state-of-the-art algorithms. To sum up, our contributions are directed towards two main aims: (i) the enhancement of the rumour detection performance on the two health related datasets with a novel solution for feature selection, and (ii) the validation of such proposal in comparison with other state-of-the-art algorithms. To this end next section briefly reports an overview on current families of feature selection algorithms.

5.1 Feature Selection techniques

As crucial step in statistical and machine learning problems, feature selection (FS) has a long history. Even if its objectives are manifold, it is possible to highlight tree main goals (Saeys et al. 2007): first, to improve model performance and prevent overfitting; second, to create fast and cost-effective models reducing the feature space dimensionality; third, to gain a deeper insight on the process that generated the data. Hence, these algorithms must search for the "best" feature subset, in terms of its capability of fulfilling one or more of the aforementioned aims.

Rosa Sicilia

In particular, referring to the classification context, FS algorithms are first divided depending on the type of task as supervised, semi-supervised and unsupervised. With respect to our proposal, we hereby analyze the taxonomies available in the literature on the supervised feature selection methods. Referring to the widely used categorization presented by Saeys et al. (2007), it is possible to distinguish FS techniques in three groups, according to how the descriptor search is combined with the classification model construction:

- *filter* methods, that evaluate the relevance of the descriptors relying only on the intrinsic properties of the data;
- *wrapper* methods, which use the model classification performance to assess the features quality;
- *embedded* methods, in which the search of the optimal set is built into the classifier construction.

Another and more detailed taxonomy of FS methods is presented in the survey offered by Huang (2015), where the feature selection algorithms are distinguished on the basis of their outcomes. In particular the authors identify three main categories:

- *feature ranking* methods, which refer to all FS techniques that provide a degree of dependency of individual features with respect to the target concepts. These methods can be further divided into pair-wise, that consider one feature at a time, and simultaneous, which consider all the features at once catching also their joint relationships;
- *subset selection* methods, that comprehend the FS techniques returning a subset of descriptors relevant to the task at hand. As further distinction there are the configuration methods, that exploit correlation among the features to identify the best subset, and the evaluation methods, which assess the subsets according to a criterion that could be model-dependent or model-independent.
- *embedded* methods: defined as before.

In Fig. 5.1 we point out the intersection between the FS categorization presented by Saeys et al. (2007) and Huang (2015), showing a more detailed association between groups of methods and the taxonomy. Nevertheless, even if each of these techniques has its advantages and downsides, the general consensus is that no feature selection method is universally superior to the others (Huang 2015).

As mentioned before, in section 4.2 we used a wrapper approach to identify the best subset, however this type of techniques are strongly dependent on the classification model used for evaluation. On the one hand, this dependency ensures greater

Rosa Sicilia

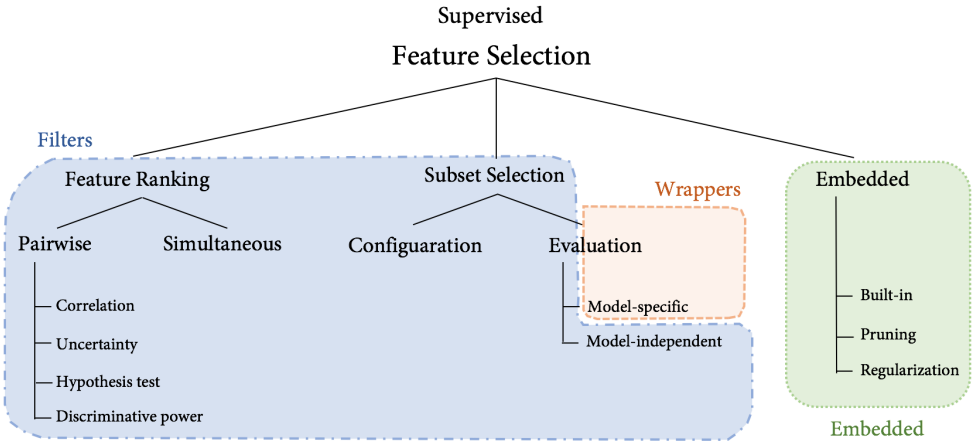


Fig. 5.1 Schematic representation of the feature selection taxonomy reported in the survey by Huang (2015). Coloured boxes identify the parallelism with the most common categorization by Saeys et al. (2007).

enhancements in performance, since the search of the descriptors takes into account the learning structure of the classification algorithm. On the other hand, this could be also a downside, since the chosen subset is specific for one classifier. Moreover, this characteristic sometimes makes wrappers prone to the overfitting issue. For this reason, as a preliminary analysis we explored the potential of a feature ranking algorithm in section 4.3, gaining promising results. This inspired us to investigate a novel feature selection algorithm that belongs to the broad category of filters, and also to provide a more comprehensive and robust analysis of feature selection for rumour detection.

5.2 Novel Approach

The general pipeline adopted in this chapter is represented in Fig. 5.2 and recalls the previous approach presented in chapter 4. It can be divided in a training and in a test phase, depicted in the upper and lower panels of Fig. 5.2, respectively. The former consists of data acquisition, network reconstruction, feature extraction, feature selection and model training. Network reconstruction recreates the conversation graphs, i.e. retweets and replies graphs, to capture the network structure and model how the news spread. Then, the second block extracts a set of hand-crafted descriptors presented in chapter 4, and, next, third block selects the most informative descriptors using a novel feature selection algorithm that we introduce in section 5.2.1; in Fig. 5.2 this part is highlighted by a red rectangle since it represents the bulk of this chapter. Finally, the best feature set is used to train the classifier. In the test phase, after again data acquisition and network reconstruction, it is possible to compute the selected features and then to test the learning model.

Rosa Sicilia

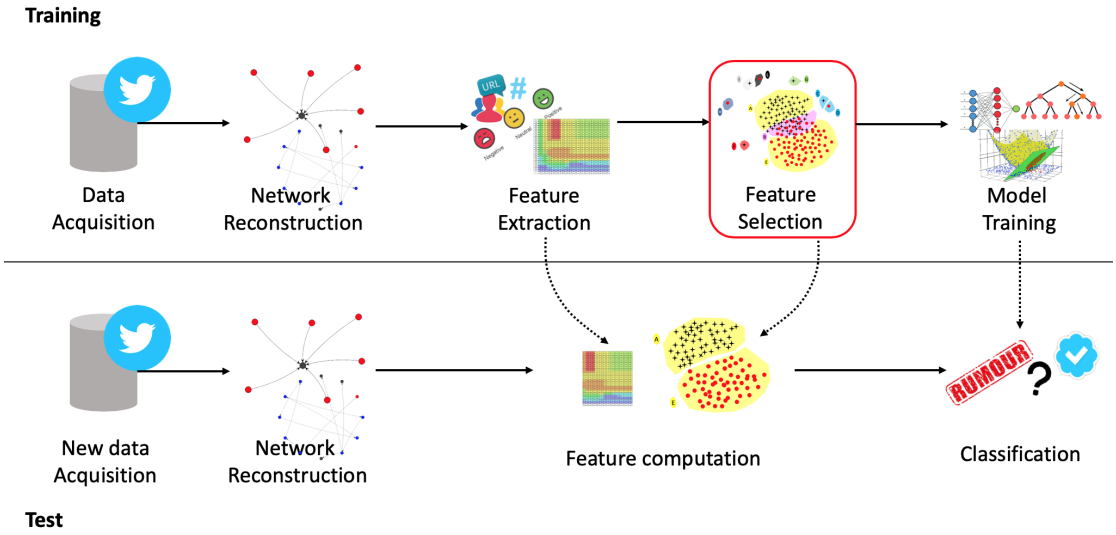


Fig. 5.2 Pipeline of the rumour detection process.

5.2.1 Proposed filter

The proposed FS method can be divided into three main steps depicted in Fig. 5.3: subset search, Rule-based Space Characterization (RSC) and subset evaluation. The first one aims at iteratively selecting specific subsets of features by exploiting the sequential backwards search (SBS), i.e. a well-known heuristic search strategy that starts from the full set of descriptors and removes one feature at a time according to a performance criterion used in the following subset evaluation stage (Huang 2015). The second block is the core of the method and it is based on a novel algorithm that aims at catching the configuration of samples in the feature space. The final step refers to the assessment criteria that is used to identify the "best" group of descriptors. Sections 5.2.2 and 5.2.3 now focus on the last two blocks constituting our main contributions.

5.2.2 Rule-based Space Characterization

As stressed so far, feature selection aims at identifying the "best" subset of descriptors for a given machine learning task and, generally, the quality of a set of features

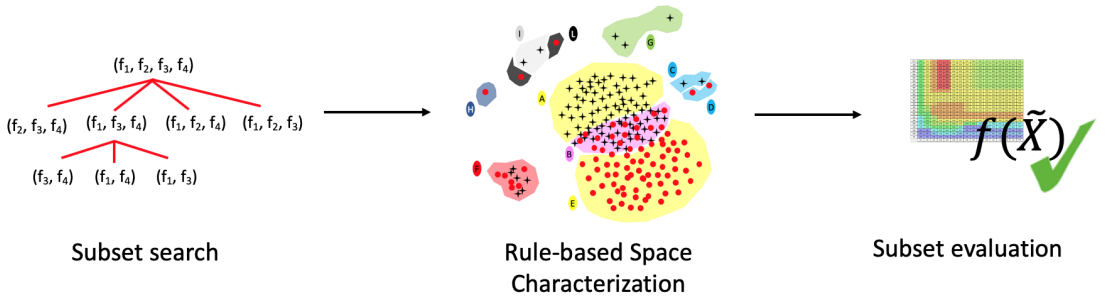


Fig. 5.3 Pipeline of the proposed feature selection approach.

Rosa Sicilia

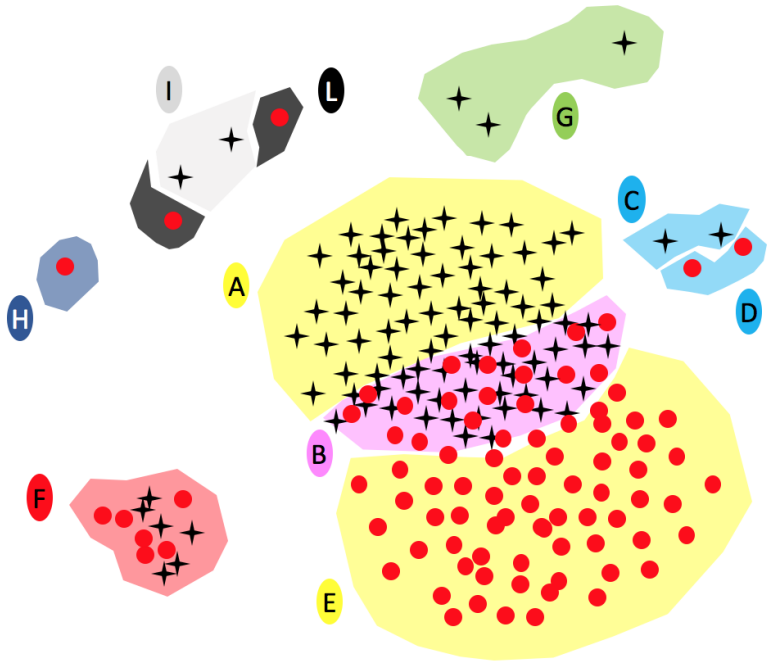


Fig. 5.4 Example of a 2-dimensional feature space in a binary classification problem with different possible arrangement of samples. Classes are distinguished by colour and shape (black stars and red circles), whereas letters and coloured areas refer to 8 configurations that can represent reliable or doubtful cases for the classification algorithm. A comprehensive explanation is reported in section 5.2.2.

is strongly related with model errors' reduction. In a classification framework, these errors may be related to particularly complex situations in the feature space which can give rise to unreliable outputs, i.e. the presence of outliers or regions with different classes overlapping. Hence, the intuitive idea behind the proposed method is that the best feature subset is the one that creates a configuration of samples that minimizes the presence of such complex and unreliable situations. For the sake of presentation, Fig. 5.4 depicts an example of 2-dimensional feature space in a binary classification problem, where different arrangements of samples are shown. This figure helps us introducing the eight possible reliable or doubtful cases for a classification algorithm.

Case A/E these cases represent main concepts of the two classes, i.e. regions with samples of the same class with no overlaps; they are referred to as clear regions in the following;

Case B the region of the feature space shows a certain degree of overlaps between samples of different classes and, hence, it is referred to as overlapped region;

Case C/D in these configurations there are samples of different classes coupled together that, furthermore, are located far from the main concept. We can consider them as coupled subconcepts of an overlapped region;

Case F this situation represents a set of instances laying far from the main concepts and

Rosa Sicilia

their nearest samples belong to an overlapped region. It can be therefore considered as a subconcept of an overlapping region;

Case G This is the case of isolated samples lying far from those belonging to a clear region of the same class; hence, this is a subconcept of a clear region;

Case H this situation represents an isolated sample that, differently from case G, is far from a clear region containing samples belonging to a different class;

Case I in this configuration the sample is far the clear region containing samples of the same class, while its closest sample is a member of the other class;

Case L this configuration is the counterpart of previous case since it refers to a sample far from a clear region containing samples of a different class, while its nearest sample is a member of the other class, being also an outlier of that class.

Except for case A/E and G, when the other configurations occur discriminative rules are hard to induce. We deem that the worst case is B since it directly hinders the separability of the samples in the feature space. Hence, finding the subset of descriptors that minimizes the number of training samples belonging to a confused region could improve classification accuracy.

We now introduce three *meta-features* that will be used to characterize the feature space in the aforementioned eight cases by means of a set of rules. Before that, we first need to introduce the concept of Tomek link and of sample neighbourhood.

A Tomek link is defined as a pair of minimally distanced nearest neighbours of opposite classes (He and Garcia 2009). Formally, let S be a dataset of m samples, such that $S = \{(\mathbf{x_i}, y_i)\}$, with $i = 1, \ldots, m$, where $\mathbf{x_i} \in X$ that is the n-dimensional feature vector of the i-th sample, whilst $y_i \in Y$ is its label. Given an instance pair $(\mathbf{x_i}, \mathbf{x_j})$, let $d(\mathbf{x_i}, \mathbf{x_j})$ be the distance measure between the two samples. When $y_i \neq y_j$, the same pair $(\mathbf{x_i}, \mathbf{x_j})$ is a Tomek link if there is no instance $\mathbf{x_k}$ such that $d(\mathbf{x_i}, \mathbf{x_k}) < d(\mathbf{x_i}, \mathbf{x_j})$ or $d(\mathbf{x_j}, \mathbf{x_k}) < d(\mathbf{x_i}, \mathbf{x_j})$. In the following, if $\mathbf{x_i}$ belongs to a Tomek link we assign 1 to the indicator variable $b_T(\mathbf{x_i})$, 0 otherwise.

Furthermore, let $I_\rho(\mathbf{x_i})$ be the neighbourhood of radius ρ of a sample $\mathbf{x_i}$, $I_\rho(\mathbf{x_i}) = \{\mathbf{x} \in X | d(\mathbf{x}, \mathbf{x_i}) \leq \rho\}$. In our experiments, we set $\rho = \max_{\forall \mathbf{x_i} \in X} \min_{\forall \mathbf{x_j} \in X} d(\mathbf{x_i}, \mathbf{x_j})$ where, straightforwardly, X represents the training set.

On these premises for each sample $\mathbf{x_i}$ we define the following meta-features:

- *Fraction of Tomek (fr_T)*: given ρ, $fr_T(\mathbf{x_i})$ for a sample $\mathbf{x_i}$ is defined as the fraction of samples belonging to a Tomek link (i.e. that have $b_T = 1$) in $I_\rho(\mathbf{x_i})$. Formally,

$$fr_T(\mathbf{x_i}|\rho) = \frac{\sum_{\mathbf{x} \in I_\rho(\mathbf{x_i})} b_T(\mathbf{x})}{|I_\rho(\mathbf{x_i})|} \tag{5.1}$$

- *number of same class samples (n_C)*: given the distance radius ρ, for a sample $\mathbf{x_i}$ with class label y_i, $n_C(\mathbf{x_i}|\rho)$ is the number of samples of the same class y_i

Rosa Sicilia

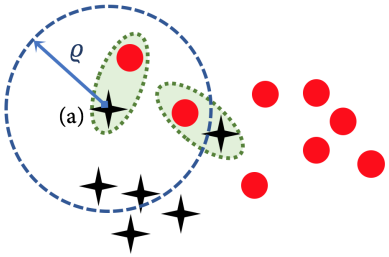


Fig. 5.5 Example of the computation of the meta-features, fr_T and fr_C for sample (a), in a 2-dimensional feature space. Shape and colour distinguish between the two classes, whereas green dotted areas point out the Tomek links.

belonging to a neighbourhood of radius ρ from $\mathbf{x_i}$. Formally,

$$n_C(\mathbf{x_i}|\rho) = \sum_{\mathbf{x} \in I_\rho(\mathbf{x_i})} b_c(\mathbf{x}, \mathbf{x_i}) \tag{5.2}$$

where $b_c(\mathbf{x}, \mathbf{x_i})$ is 1 if $\mathbf{x}$ belongs to the same class of $\mathbf{x_i}$, and 0 otherwise.

- *Fraction of same class samples (fr_C)*: similarly to $n_C(\mathbf{x_i}|\rho)$, given the distance radius ρ, for a sample $\mathbf{x_i}$ with class label y_i, $fr_C(\mathbf{x_i})$ is the fraction of samples of the same class y_i belonging to a neighbourhood of radius ρ from $\mathbf{x_i}$. Formally,

$$fr_C(\mathbf{x_i}|\rho) = \frac{n_C(\mathbf{x_i}|\rho)}{|I_\rho(\mathbf{x_i})|} \tag{5.3}$$

To better understand these concepts, Fig. 5.5 offers a graphical example of two set of samples of different classes distinguished according to shape and colour in a 2-dimensional feature space. Given the sample (a) and its neighbourhood of radius ρ, the computed meta-features would be $fr_T = 0.5$ because there are two Tomek links (denoted by green dotted areas) and four samples in its neighbourhood; $n_C = 2$ since in its neighbourhood two samples belong to the same class of (a); $fr_C = n_C/4 = 0.5$ four samples are in its neighbourhood. From these measures it is possible to grasp information about an instance and its neighbourhood: for instance, a high fr_T score could indicate that the region under consideration is an overlapped region, whilst a high fr_C could point out the presence of a clear region.

On these grounds, we design a heuristic rule-based decision tree distinguishing among the eight cases listed before, which are based on different combinations of the three meta-features. Fig. 5.6 represents the tree where the rules are highlighted in bold red. We modelled values that tend to 0 or 1 (i.e. $\rightarrow$) with two parameters, namely ε and β, that are used as lower and upper thresholds, respectively. Hence, if a meta-feature $f \rightarrow 0$ it will be modelled as $f \leq \varepsilon$, whereas if $f \rightarrow 1$ it will be considered as $f \geq \beta$. To be consistent, if a meta-feature is $0 < f < 1$, it is implemented as $\varepsilon \leq f \leq \beta$.

Rosa Sicilia

Furthermore, in the figure $\theta \in [0, 1]$, and N is the total number of samples belonging to one of the classes in Y.

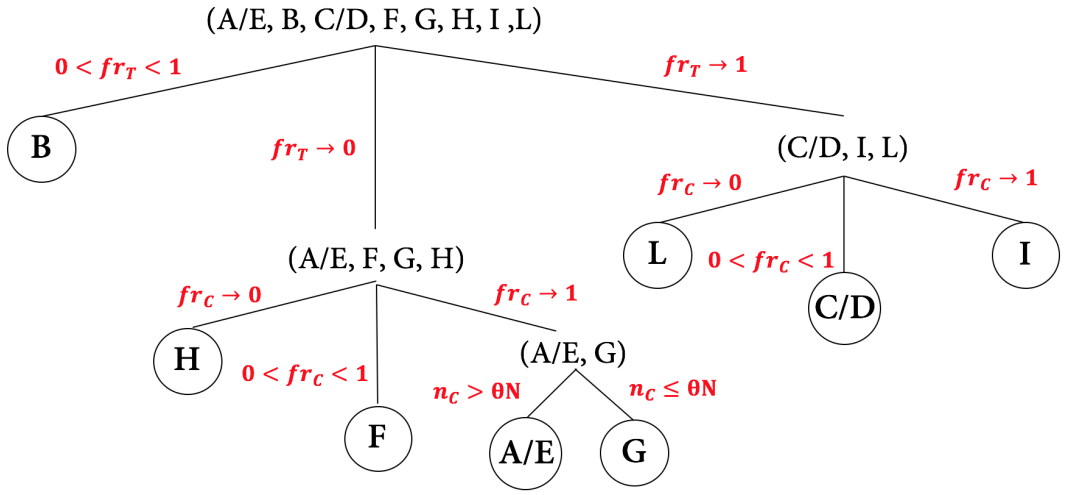


Fig. 5.6 Rule-based decision tree presenting how the training samples are assigned to one of the eight classes using the meta-features.

5.2.3 Feature subset evaluation criteria

In the sequential backward search, at each step there is the need to evaluate different feature subsets to find out the best one. To this goal, the proposed method minimizes a criterion function that weights the samples on the basis of case label assigned by the RSC tree. Formally, let $\tilde{X}$ be a subset of a given feature space X ($\tilde{X} \subseteq X$), with n being the dimension of X and m the dimension of $\tilde{X}$, such that $m \leq n$. We define a 1×8 vector $\mathbf{z}(\tilde{X})$, where each component $z_j(\tilde{X})$ is equal to the following expression:

$$z_j(\tilde{X}) = \sum_{\tilde{\mathbf{x}}_i \in \tilde{X}} RSC_j(\tilde{\mathbf{x}}_i) \tag{5.4}$$

where $RSC_j(\tilde{\mathbf{x}}_i)$ is the output of the Rule-based Space Characterization algorithm for the sample $\tilde{\mathbf{x}}_i \in \tilde{X}$. It returns 1 if $\tilde{\mathbf{x}}_i \in$ class j, 0 otherwise. In other words, vector $\mathbf{z}(\tilde{X})$ contains in position j the count of samples belonging to the jth case of the RSC algorithm, run with the specific subset of features $\tilde{X}$. Hence, the criterion function is defined by the following equation:

$$\min_{\forall \tilde{X} \subseteq X} f(\tilde{X}) = \frac{\mathbf{w}^{\mathbf{T}} \mathbf{z}(\tilde{X})}{\|\mathbf{w}\| \|\mathbf{z}(\tilde{X})\|} \tag{5.5}$$

where $\mathbf{w} = \{\omega_j\}_{j=1}^{8}$ is a 1×8 vector of weights. It is worth specifying that assigning a grater weight w_j to a specific class j implies that the feature selection method

Rosa Sicilia

would return as best feature subset the one that minimizes the number of samples with configuration j.

With reference to the FS taxonomy shown in Fig. 5.1, our proposal can be categorized into the group of model-independent subset evaluators for the following three reasons: (i) it belongs to subset selection category since it outputs the best subset of features; (ii) it uses an evaluation criterion rather than any correlation measure and, hence, it can be placed in the evaluation group; (iii) it fits in the Model-independent class since the assessment of subsets is independent from any specific learning model because it is based only on the characteristics of the data.

5.3 Experimental Setup

The proposed method was tested on the datasets described in chapter 3. All the experiments, accounting feature selection and system performance evaluation, were run in a 10-fold cross validation on both datasets (#Zikavirus and #Vaccine). Furthermore, the use of nested cross validation avoided any bias between feature selection phase and performance evaluation. We also considered classifiers belonging to different learning paradigms, including a Multi-Layer Perceptron (MLP) as a neural network, a k-Nearest Neighbour (kNN) as an instance-based classifier, a Support Vector Machine (SVM) as a kernel machine and a Random Forest (RF) as a classification ensemble. For all of them the parameters were set to the default values, even if we acknowledge that their tuning could lead to better results. Nevertheless, the No Free Lunch Theorems for Optimization tell us that all configurations perform equally well when averaged over all possible experiments (Wolpert and Macready 1997). As basis for comparison, we therefore preferred to maintain the baseline parameter set, so that the only way a method can outperform another one relies upon its specialization to the given feature set (Ho and Pepyne 2002).

The proposed method has a total of three parameters, i.e. β, ε and θ for space characterization. After preliminary experiments, we found that $\theta = 0.02$ is a reasonable value, and its variation does not affect the results. For this reason in the feature selection stage we tune only β and ε in $(0.6, 1.0)$, and in $(0, 0.5)$, respectively. These parameters are searched to determine which configuration attains the largest classification performance; despite that, it is worth pointing out that the whole evaluation criteria for the best subset is still model independent. Furthermore, the weight vector $\mathbf{w}$ used in the evaluation criteria, introduced in section 5.2.3, should have grater values for those configurations that could result in classification errors. We deem that the worst case is related to overlapped regions, represented by configuration B in Fig.5.4, where there is a mixture of samples belonging to different classes. Hence, we set to 1 weight for case B, whereas we assigned 0 to all others. This choice ensures that the proposed feature

Rosa Sicilia

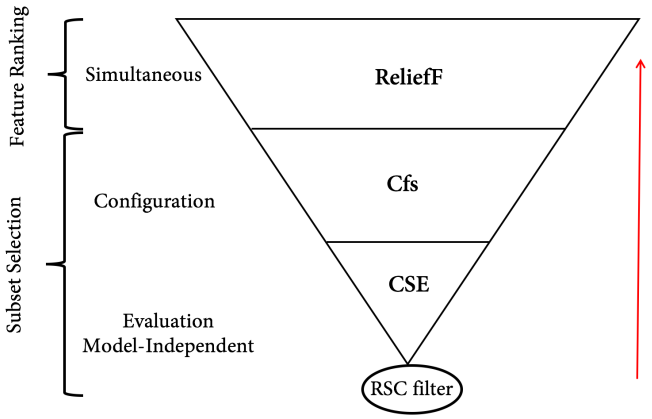


Fig. 5.7 Bottom-up path for competitors identification. The reverse pyramid starts from the proposed method (RSC filter) and at each level shows the corresponding competitor. As we go further from our method the base enlarges, as the difference with respect to the RSC filter increases.

selection algorithm would have as primary focus the search of the best subset reducing the presence of overlapped regions for the classifier.

At the end of previous section we discussed how our method establishes with respect to the taxonomies adopted in the literature. This analysis also permits us to select the feature selection methods engaged as competitors in our experiments. They are three FS filters having different degrees of similarity with our proposal, according to the categorization in Huang (2015). Figure 5.7 depicts a reversed pyramid representing a bottom-up path along the taxonomy tree (Fig.5.1). It starts from the proposed RSC filter and on each level reports a different FS method belonging to the corresponding category of the tree. Moving in a bottom-up fashion along the taxonomy, we start from the Model-independent Subset Evaluation methods, choosing as first competitor the Consistency Subset Evaluation (CSE). This method evaluates the worth of a subset of attributes measuring the consistency in the class values when the training instances are projected onto the subset of attributes (Liu et al. 1996). Proceeding one level up in the tree we find the Configuration Subset selection algorithms, from which we selected the Correlation-based feature selection (Cfs) method as a second competitor. Cfs assesses a feature subset exploiting a correlation measure to estimate the individual predictive ability of each feature along with the degree of redundancy between them (Hall 1999). Taking a final step up along the tree, we chose as last competitor a Simultaneous Feature Ranking approach, i.e. the ReliefF. It computes relevance weights for all the features at the same time by looking into their joint relationship with the target concept (Kononenko 1994). Since the latter is a ranking method, we used a threshold-based selection step to obtain the best subset. To run the experiments applying these FS approaches we used their standard implementation available within the WEKA framework (Witten et al. 2016).

Rosa Sicilia

Feature selection	***#Zikavirus***				***#Vaccine***			
	kNN	**SVM**	**RF**	**MLP**	**kNN**	**SVM**	**RF**	**MLP**
RSC	**.824**	**.761**	**.874**	**.781**	**.799**	.964	**.968**	.659
CSE	.770	.756	**.874**	.756	.570	.564	.960	.625
Cfs	.822	.757	.870	.770	.798	**.967**	**.968**	.610
ReliefF	.802	**.761**	.869	.771	.797	**.967**	.96	**.665**

Table 5.1 Best accuracy achieved by the rumour detection system using different FS and classification approaches on the two datasets. In bold the highest accuracy values are emphasized.

In next section we will present and discuss the results obtained with the aforementioned setups.

5.4 Results

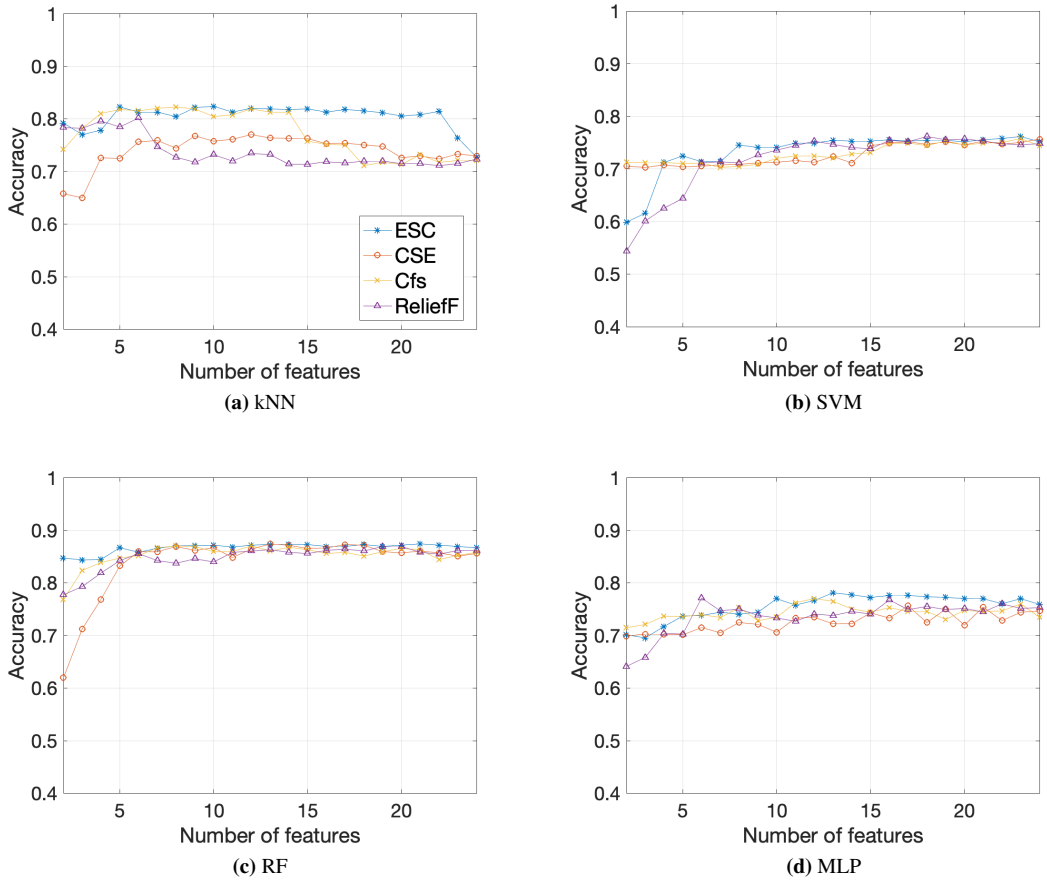
Table 5.1 presents the best performance of the rumour detection system, combining all feature selection algorithms with the selected pool of classifiers described in section 5.3. For the sake of brevity, we report only the mean accuracy on the cross validation folds, since on average all the other performance metrics (e.g. recall, precision, f-score) followed the same trend.

In Table 5.1 the row named as RSC shows the results achieved using the proposed FS method: it is worth noting that it outperforms other methods in most of the cases. In particular, in comparison with the CSE algorithm, that is the method that belongs to the most similar category according to Fig. 5.1, RSC performance is always higher regardless of the dataset and of the learning paradigm adopted, except for one tie. Moreover, it is also interesting to compare the accuracies attained with those already reported in the previous chapter in section 4.3[1], with an accuracy upgrade of $\sim$1% and $\sim$5% on the #Vaccine and the #Zikavirus datasets, respectively. This suggest us that the rumour detection system benefits from the proposed feature selection approach.

To get a deeper insight in the benchmark with competitors, we analysed how performance varies with the number of features selected by the filters. In this respect, the four plots in Fig. 5.8 show the accuracy on the #Zikavirus dataset for the four classification algorithms and for all the tested feature selection methods, distinguished by different colours and markers shape. Each plot reports on x-axis the number of features selected by each method and on y-axis the accuracy. As an overall consideration, it is possible to notice that the blue line corresponding to the RSC performance reaches higher accuracy in most of the cases. Although these trends do not show evident fluctuations,

[1]We refer to section 4.3 and not to 4.2.2 since the performance reported in the former is higher than the one presented in the latter.

Rosa Sicilia



(a) kNN **(b)** SVM **(c)** RF **(d)** MLP

Fig. 5.8 Average accuracy trends on the #Zikavirus dataset considering different subsets of features. Each panel corresponds to a learning paradigm. Each curve reports the performance achieved by the detection system using a specific filter for FS, and the legend shown in panel (a) applies to all the other panels.

we note that the proposed method exhibits a slight drop in performance with the kNN architecture for subsets with a great number of features ($\geq$20). We deem that this phenomenon could be related to a slight curse of dimensionality caused by the presence of unnecessary features that hinder the classification capabilities on this dataset of the specific learning algorithm.

Fig. 5.9 shows the accuracy variations on the #Vaccine dataset, and it reveals a similar trend. However, in this case the performance drop is large for both SVM and kNN. On the one hand, in the case of the kNN this could be given again by the curse of dimensionality. On the other hand, since the SVM classifier is less prone to this issue, the performance drop observed for this classifier could be due to the features included in the last subset that could deteriorate the classification ability of the set of descriptors. Furthermore, it is worth also noticing that RSC filter displays the accuracy drop at a number of features larger than the other methods, proving its capability to find subsets that provide discriminant configurations in the feature space. The results show

Rosa Sicilia

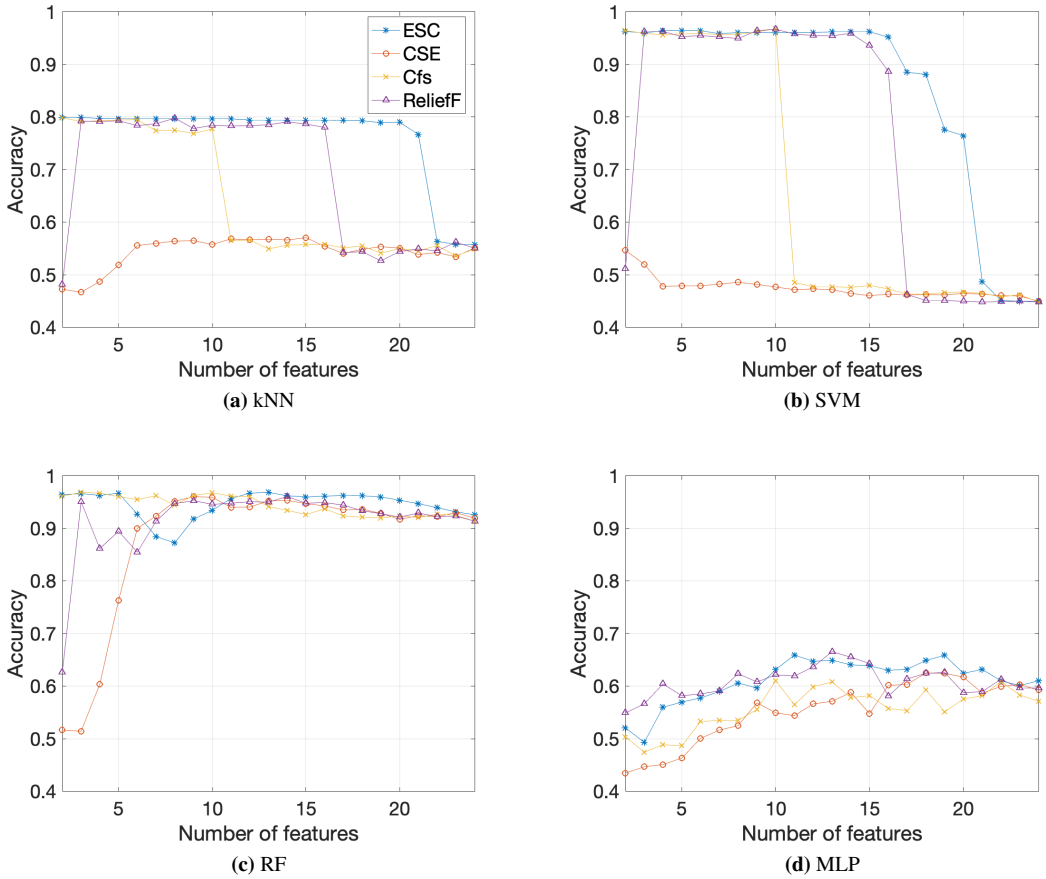
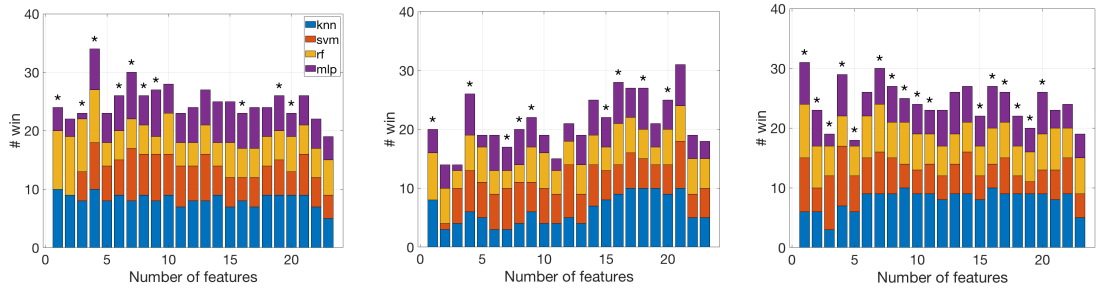


Fig. 5.9 Average accuracy trends on the #Vaccine dataset considering different subsets of features. Each panel corresponds to a learning paradigm. Each curve reports the performance achieved by the detection system using a specific filter for FS, and the legend shown in panel (a) applies to all the other panels.

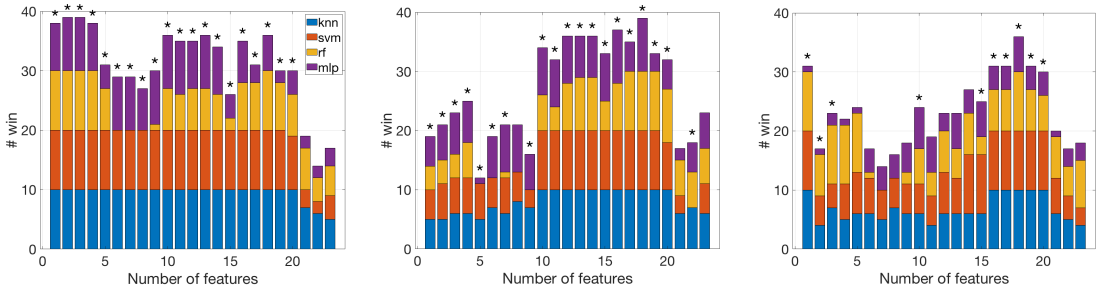
that the MLP attains lower accuracy along with slight fluctuations in both trends. We deem that could be due to the default number of neurons in the hidden layer that, as mentioned in section 5.3, was not optimized for the specific problem. Nevertheless, for the motivations already discussed there, we preferred to maintain the baseline parameter set.

To further compare the tested filters we perform a pairwise analysis computing win-tie-loss scores of our filter with respect to the others. For each number of selected features, and given a learning paradigm, we pairwise compare the accuracies achieved on each test fold of the cross validation by two feature selection methods. Hence, the maximum number of wins is 10 for each classifier and, summing up the values for all the learners, the number of wins range from 0 to 40. Each of the three panels in Fig. 5.10 shows the stacked histograms reporting the amount of wins when we compared the RSC filter against the competitors. Using the Wilcoxon rank sum test, we also computed the p-value of the pairwise difference between the performance of the



(a) RSC vs CSE **(b)** RSC vs Cfs **(c)** RSC vs RF

Fig. 5.10 Stacked histograms of the number of wins of the proposed method against the competitors computed on the #Zikavirus dataset. Each bar shows in different colours the contribution of each classifier combined with the RSC filter. The legend shown in panel (a) applies to all the other panels.



(a) RSC vs CSE **(b)** RSC vs Cfs **(c)** RSC vs RF

Fig. 5.11 Stacked histograms of the number of wins of the proposed method against the competitors computed on the #Vaccine dataset. Each bar shows in different colours the contribution of each classifier combined with the RSC filter. The legend shown in panel (a) applies to all the other panels.

feature selection methods. The results are reported again in Fig. 5.10, where a star (*) above the stacked histogram bars means that we found performance differences that are statistically significant ($p \geq 0.1$).

All the three panels show that RSC combined with kNN outperforms the competitor, whereas the other classifiers present slight fluctuations depending on the feature subset. Turning our attention to the number of wins in the case of the #Vaccine dataset, shown in Fig. 5.11, we observe that RCS wins to a large extent when combined with both the kNN and the SVM classifiers. Moreover, on #Vaccine dataset RSC achieves a statistical significant win in almost all configurations with respect to the CSE, which is its nearest neighbour in the taxonomy (Fig.5.1).

5.5 Remarks

In this chapter we presented a novel method for feature selection, which is a model-independent subset evaluation filter. It was applied to detect rumours on two health-

Rosa Sicilia

related Twitter datasets. With respect to the previous chapter, where we presented a hand-crafted features set with newly designed descriptors (sec. 4.1) and we analysed the robustness of this set on different health topics (sec. 4.3), this contribution focuses on the feature selection procedure as a key step for this machine learning application, and in general for any other ones. The proposed method relies on a new Rule-based Space Characterization algorithm that aims at identifying samples lying in complex and less reliable regions in the feature space. It selects the feature subset reducing the number of samples in those regions. Such feature selection procedure allows to understand which are the descriptors that provide separable configurations in the feature space and to reach those configurations. So with further analysis this method could potentially give information about the reliability of the classification. Broadly speaking, the proposed approach can be viewed as a model for explanatory artificial intelligence (XAI) since it is able to summarize the reasons for classifier behaviour producing insights about the causes of their decisions (Gilpin et al. 2018).

To validate our approach, we compared it with other feature selection methods that belong from the same group, according to well-established taxonomies in the literature. Results on two Twitter health datasets showed that in most of the combinations our proposal outperforms the other filters. Best performance was obtained on both datasets using our proposal for feature selection and a Random Forest for classification. Moreover, we were able to boost the accuracy on both the datasets tested.

This findings prove the potential of the RSC filter, nonetheless they also bring a step forward the micro-level rumour detection analysis.

Rosa Sicilia

CHAPTER 6

Transferring Knowledge across topics

The analyses presented so far have focused on a rumour detection system able to classify Twitter posts of a specific topic domain. The research question we bring up in this chapter is: *How does the system perform when training and test samples belong to different topics?* Answering this question has two main consequences: on the one hand, if the system was robust to cross-topic tests, it could potentially be applied to detect newly rising rumours that usually belong to unknown topics in the training set; on the other hand, if the system presented so far were not able to discriminate rumours in cross-topic environments, a different strategy for knowledge transferring would be necessary to enhance system performance in a potential unknown environment. Straightforwardly, to be used in a real-life environment any system trained on a set of rumour post should be sufficiently able to generalize to rumours of yet unknown topics, which are rising at the moment. This is not trivial since different rumour domains can have intrinsically different characteristics while spreading given also the different people reactions they can induce, and these differences could hinder the system performance.

In the following section we will first present a preliminary analysis of cross-topic tests on our two health related datasets (sec. 6.1), then we will introduce the transfer learning techniques (sec. 6.2) and their application to our case of study, proposing a novel method (sec. 6.3-6.7).

6.1 Preliminary Cross-topic analysis

In this exploratory study we considered the same general pipeline presented in chapter 4. For the sake of clarity, in figure Fig. 6.1 we recall a simplified scheme of the process: the upper panel shows the training process, whereas the lower panel represents the test procedure. The samples retrieved are first analysed to reconstruct the conversation graphs and the network structure. Then we extract our set of descriptors presented in section 4.1 selecting the best features for the problem at hand. Finally, this best feature set is used for classifier training and then for the test phase on unknown samples.

In the following sections we briefly present the structure of the experiments performed.

Rosa Sicilia

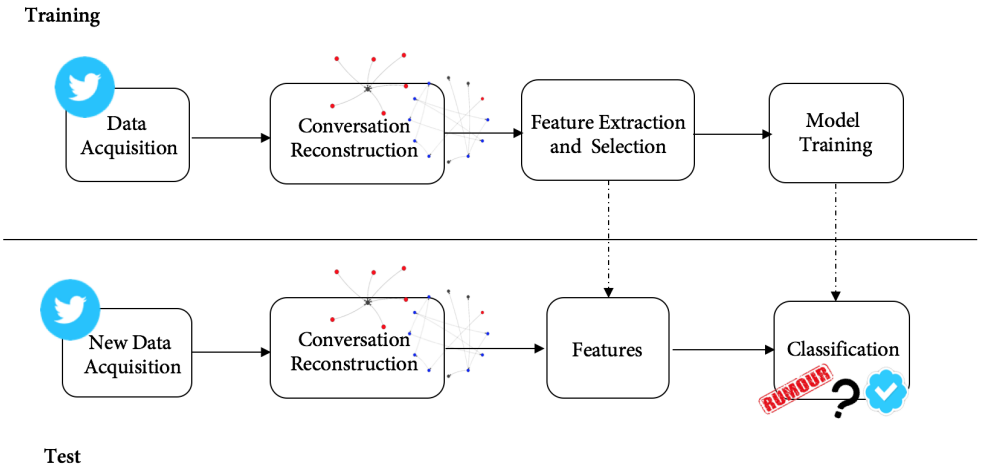


Fig. 6.1 Pipeline of the rumour detection process.

6.1.1 Experimental Design

In order to study the behaviour of the system in cross-topic tests we explored the variation in the prediction of the system when trained on a fixed dataset (hereinafter called training topic) and tested on the another one (hereinafter called test topic). In particular, referring to Fig.6.2, we analysed how much the performance changes when, fixed the samples of training topic, we vary the amount of training samples belonging to the test topic (straightforwardly, these training samples belonging to the test topic are not the same used in the test set). To this end we divided the test topic into 8 folds, containing at least 100 samples, and at each iteration one fold was used for actual test, whilst the others were used to progressively increment the size of the training set. With this setup we performed two experiments:

EXP#1 In this experiment the training topic was #Zikavirus and the test topic was #Vaccine .

EXP#2 In this second experiment we swithced the dataset, i.e. #Vaccine was used as fixed training topic, whereas #Zikavirus was used as test topic in the aforementioned incremental fashion.

6.1.2 Results

The two experiments presented in section 6.1.1 were based on the system in Fig. 6.1, skipping the feature selection step and using the best subsets found in the single topic analysis of section 4.3.2. Recalling the previous analysis, the feature evaluation and selection was performed using ReliefF, whereas the classification step employed the Random Forest ensemble as learning paradigm.

Figures 6.3a and 6.3b show the results obtained in the first experiment (EXP#1), whereas figures 6.3c and 6.3d report the performance obtained in the second experiment

Rosa Sicilia

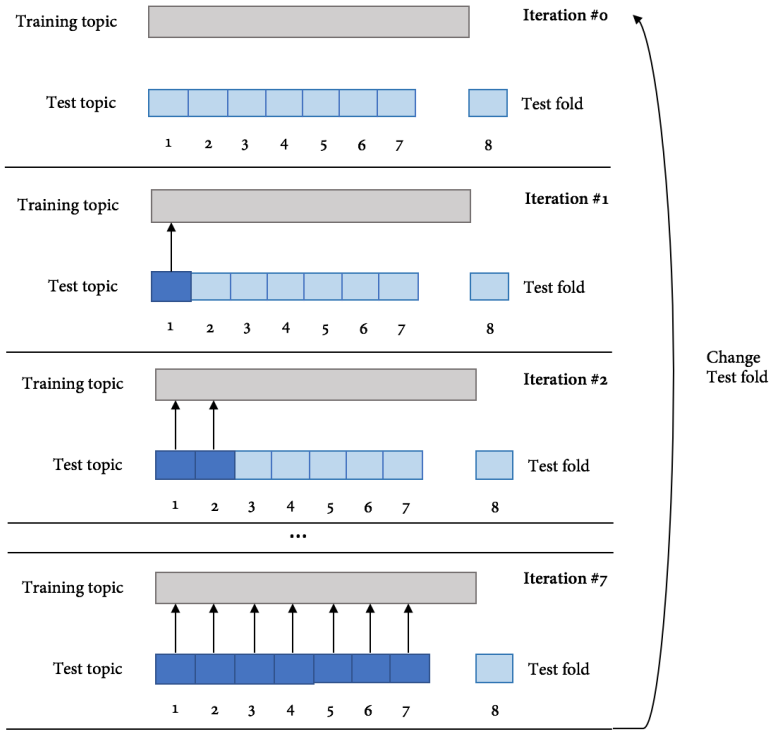


Fig. 6.2 The process followed in experiments EXP#1 and EXP#2: the test topic is divided into 8 folds, one is chosen as test and the others are progressively added to the fixed training set. The whole process is then iterated changing each time the test fold.

(EXP#2). In each plot, the y-axis reports one performance measure, whilst the x-axis shows the number of folds of the test topic that were added to the training set, ranging from 0 to 7 as mentioned in section 6.1.1. On the one side, 0 means that the training set includes only samples belonging to the training topic (different from the test one). On the other side, 7 means that the training set contains samples belonging to the training topic and samples coming from 7 out of 8 folds of the test topic. Moreover, each point in the chart represents the mean and the standard deviation of the performance index.

In both experiments, adding knowledge on the test topic in the training set significantly increases the performance of the system. Figures 6.3a and 6.3c show that the system yields to an overall accuracy under 30%, when trained exclusively on one topic and tested on the other. This indicates that even topics belonging to the same domain (health) are characterized by different rumour propagation structures and users dynamics. It is worth noticing that when using the #Zikavirus dataset as training (Fig.6.3a), the accuracy has a faster raise at the beginning, but then curve doesn't even reach 90%. A slightly different behaviour is shown for #Vaccine (Fig.6.3c), where the curve is nearly linear and the final accuracy is 96.0%. So using #Vaccine and part of #Zikavirus for training increases performance of the system for rumour detection on posts about the zika virus, which raises from 82.3%, as found in table 4.5 of section 4.3.2, to 96.02% (fig. 6.3c). On the other hand, the same consideration is inverted for rumour detection

Rosa Sicilia

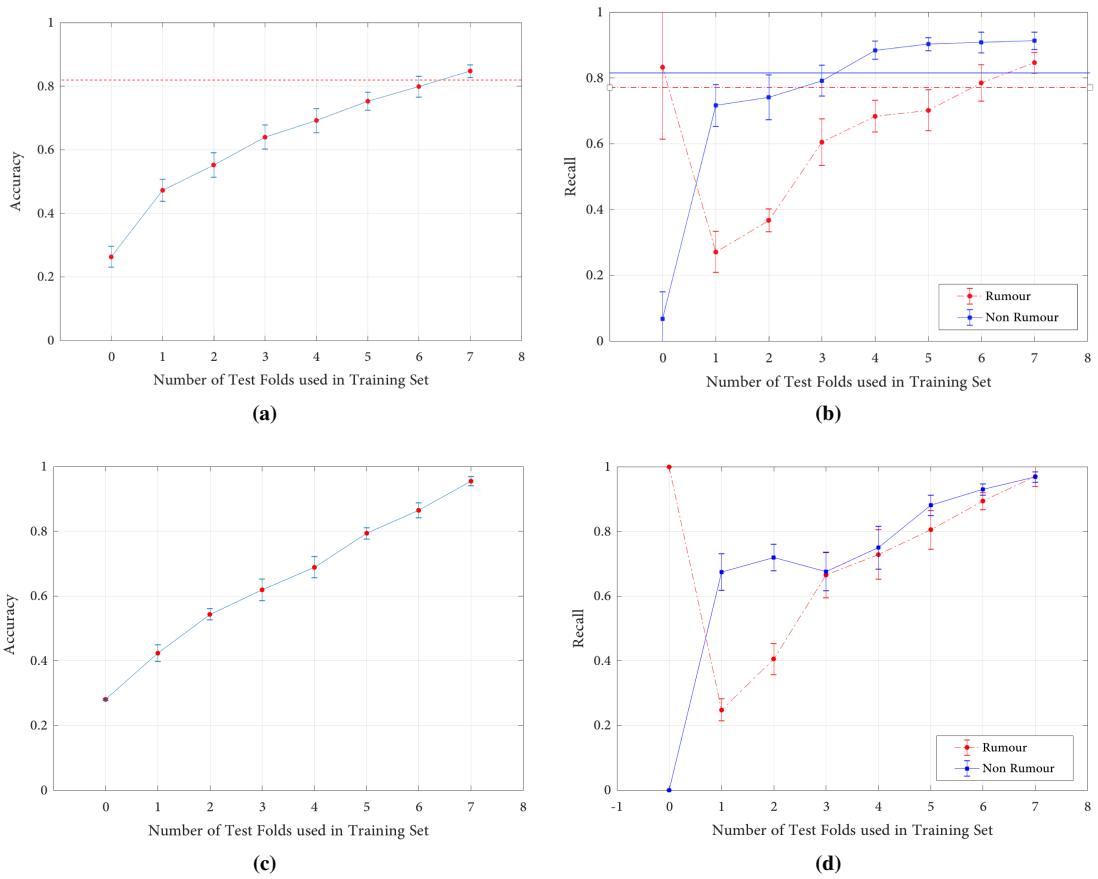


Fig. 6.3 (a) Average accuracy per number of test topic folds used in the training set for EXP#1. The dotted straight line represents the average human accuracy in rumour detection on #Vaccine dataset. (b) Average recall for rumour (dotted line) and non-rumour classes plotted against the number of test topic folds used in the training set for EXP#1. The two straight lines represents the average human recall in rumour (dotted line) and non-rumour classification on #Vaccine dataset. (c) Average accuracy per number of test topic folds used in the training set for EXP#2. (d) Average recall for rumour (dotted line) and non-rumour classes plotted against the number of test topic folds used in the training set for EXP#2.

on posts about vaccines, where the overall accuracy drops from 96.0% (section 4.3.2, tab. 4.5) to 84.1% (fig. 6.3a). Similar considerations hold also for the recall parameter shown in figures 6.3b and 6.3d, where there is only an initial error of the system that classifies all non-rumour posts as rumours. Moreover, looking at the horizontal lines in figures 6.3a and 6.3b, which represent the human performance on #Vaccine data, it is possible to notice that our system needs all 7 test topic folds to outperform a human annotator.

6.1.3 Discussion

The analysis presented so far revealed that adding knowledge of the test topic in the training set significantly increases the classification performance. However, good re-

Rosa Sicilia

sults were obtained only if there was sufficient knowledge on the test topic in the training set, meaning that at least the 80% of the test topic samples should be included in the training. We deem there are two main reasons for this findings: first, the features actually exhibit a distribution specific not only to a broad domain (health) but also to the single topic at hand (#Zikavirus or #Vaccine); second, there could be a bias introduced by the different annotation processes followed for the two dataset.

To answer the research question, this preliminary analysis leads us to explore Transfer Learning techniques to enhance system performance on different topics. To this end, next section will briefly analyze the taxonomy of state-of-the art approaches and then we will introduce a novel approach based on the Rule-based Space Characterization presented in chapter 5.

6.2 A brief overview on Transfer Learning techniques

A fundamental assumption in traditional machine learning is that training and test samples are drawn from the same domain, meaning that the data distribution and the input feature space are the same. This advantageous assumption does not hold in some real-world scenarios: indeed, many times training data is hard to collect or expensive to label. Having domain differences in training and test data could have a significant impact on classification performance, which can be considerably degraded.

Transfer Learning is the branch of machine learning that aims at studying algorithms to transfer knowledge between different domains. These techniques pursue the goal of creating systems robust to changes in data distributions or in feature spaces. In order to formally describe these concepts and to introduce the taxonomy of these techniques we exploited the notation and the definitions reported in the comprehensive survey by Weiss et al. (2016). A given domain $\mathcal{D}$ can be characterized by a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, with $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\} \in \mathcal{X}$. Moreover, it is also possible to define a task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, associated with $\mathcal{D}$, where $\mathcal{Y}$ is a label space and $f(\cdot)$ is a predictive function learned from the pairs $\{\mathbf{x_i}, \mathbf{y_i}\}$ of feature vectors, $\mathbf{x_i} \in X$, and labels, $\mathbf{y_i} \in \mathcal{Y}$. Now, given a source domain $\mathcal{D}_\mathcal{S}$ and a target domain $\mathcal{D}_\mathcal{T}$ with the corresponding source task $\mathcal{T}_\mathcal{S}$ and target task $\mathcal{T}_\mathcal{T}$, transfer learning aims at improving the target predictive function $f_T(\cdot)$ by exploiting the related information from $\mathcal{D}_\mathcal{S}$ and $\mathcal{T}_\mathcal{S}$, where $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$ or $\mathcal{T}_\mathcal{S} \neq \mathcal{T}_\mathcal{T}$. On the one hand, since we defined $\mathcal{D} = \{\mathcal{X}, P(X)\}$, the first inequality $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$ means that the feature spaces are different, $\mathcal{X}_\mathcal{S} \neq \mathcal{X}_\mathcal{T}$, and/or that there is a distinction in the marginal probabilities, $P(X_S) \neq P(X_T)$. On the other hand, from the definition of the task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, $\mathcal{T}_\mathcal{S} \neq \mathcal{T}_\mathcal{T}$ implies a mismatch in the labels space, $\mathcal{Y}_\mathcal{S} \neq \mathcal{Y}_\mathcal{T}$, and/or that the conditional probabilities are different, $P(Y_S|X_S) \neq P(Y_T|X_T)$.

Fig. 6.4 shows a schematic representation of the taxonomy presented in Weiss et al.

Rosa Sicilia

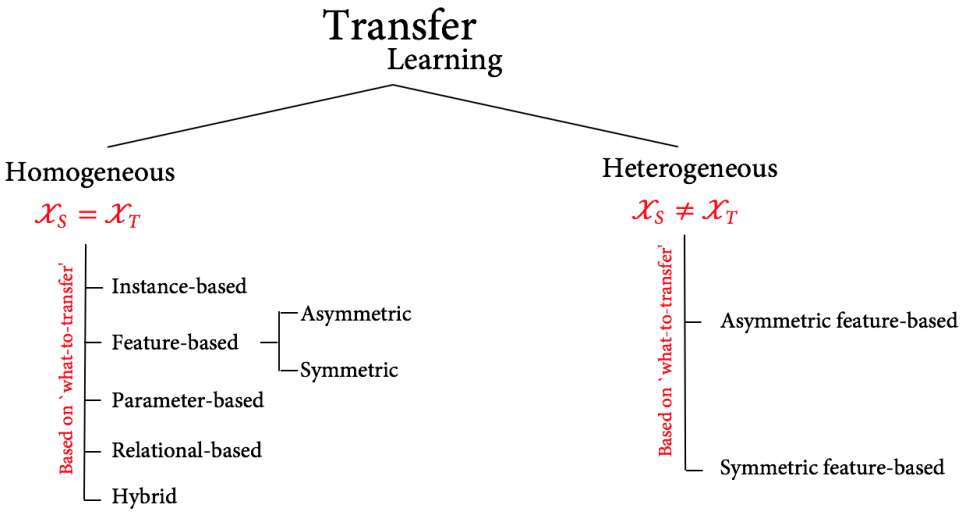


Fig. 6.4 Taxonomy of the Transfer Learning techniques based on the survey by Weiss et al. (2016). Bold red writings highlight the two main categorization rules: the first level refers to the difference in feature spaces, discriminating between homogeneous and heterogeneous transfer, whereas the second level is based on 'what-to-transfer'.

(2016), where it is possible to identify two main groups of techniques: *homogeneous* and *heterogeneous*.

Homogeneous transfer learning refers to all those methods that cope with a transfer learning problem where the source and the target domains are represented in the same feature space, i.e. $\mathcal{X}_S = \mathcal{X}_T$. In this case, most of the solutions address one of the following problems: (i) to correct the marginal distribution difference between source and target domains ($P(X_S) \neq P(X_T)$), (ii) to adjust the conditional distributions differences ($P(Y_S|X_S) \neq P(Y_T|X_T)$), (iii) to correct both. To these ends, it is possible to further distinguish homogeneous transfer learning according to 'what-to-transfer', that means the type of information that is transferred among domains:

- *instance-based*: it refers to transfer learning through instances. A common technique is to re-weight the source instances to correct the marginal distribution differences. However, this is just an example and doesn't leave out those methods that operate also on the conditional distributions.

- *feature-based*: this is the transfer learning using features. It can be distinguished in *asymmetric* feature transformation and *symmetric* feature transformation. The first approach transforms the source domain features, adjusting them to be more similar to the target. The second approach transforms both source and target features to find underlying meaningful structures, building a common latent feature space.

- *parameter-based*: this identifies the pool of techniques that transfers knowledge

Rosa Sicilia

through parameters. For instance, they commonly use shared parameters between source and target domain learning models or exploit multiple learning models.

- *relational-based*: this techniques base the knowledge transferring on some defined relationships between the source and the target domain. According to Weiss et al. (2016), this is the least used approach.

Beyond these four transfer categories, the literature enumerates also *hybrid* approaches that use more than one method to transfer knowledge among domains.

For what concerns the heterogeneous transfer learning, this category includes all those methodologies that tackle transfer problems with different source and target feature spaces ($\mathcal{X}_S \neq \mathcal{X}_T$). These methods are generally focused on aligning source and target feature spaces with the implicit or explicit assumption that the domain distributions are the same. On these grounds, it is possible to distinguish only two broad subcategories that transfer feature knowledge, namely *symmetric* and *asymmetric* feature transfer, which are similar to those previously described for homogeneous transfer learning.

6.3 Proposed method

Based on the definitions reported so far, it is possible to formalize our transfer learning problem among two domains, i.e. #Vaccine and #Zikavirus. Specifically these two domains are related as parts of the high-level "health" topic. Moreover, they are represented with the same feature space, so our solution is constrained to the homogeneous scenario. Fig. 6.5 shows a schematic representation of the general pipeline of the proposed method. First, given two datasets belonging to a source domain $\mathcal{D}_S$ and to a target domain $\mathcal{D}_T$, respectively, we align the samples in the feature space operating on the descriptor values. After that, training instances are re-weighted according to their configuration determined with the Rule-based Characterization algorithm, presented in chapter 5. Finally, the aligned and re-weighted training set is used to learn a classification model that is tested on the target domain[1]. In the followings we will fully describe each step.

6.3.1 Domain Alignment

The first block of our proposal focuses on aligning domain samples in the feature space, therefore transferring knowledge through features. Indeed, it is not guaranteed that sample clusters in source and target domains are located in the same zone of the feature space. To cope with this eventuality, we applied a rigid shift of the source samples to

[1]It is worth specifying that the target samples used in the test phase are always different from those used during training.

Rosa Sicilia

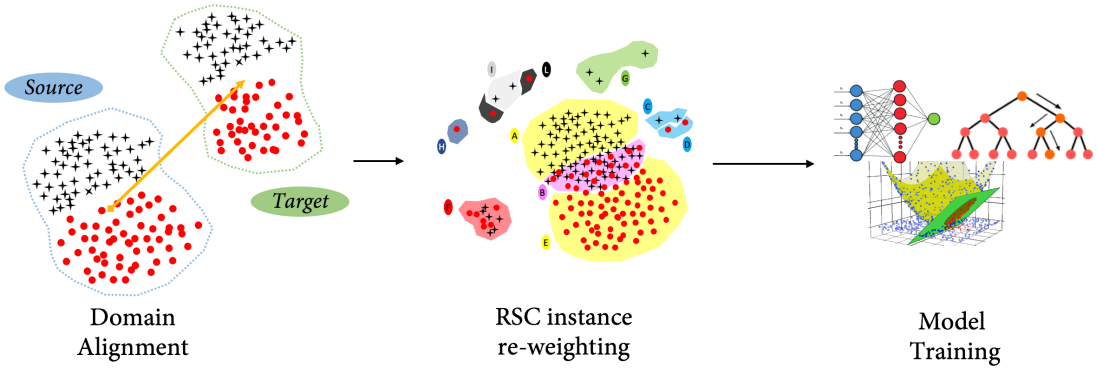


Fig. 6.5 Proposed method pipeline.

match the target set. We defined two rigid alignment types, i.e. a "global" shift and a "per class" shift.

Given a generic set of samples $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, with sample $\mathbf{x}_i \in \mathcal{X}$ and label $y_i \in \mathcal{Y}$, we can define the centroid vector $\mathbf{C}$:

$$C_j = \frac{\sum_{\mathbf{x}_i \in \mathcal{X}} x_{ij}}{N} \tag{6.1}$$

where C_j is the j-th component of vector $\mathbf{C}$, x_{ij} is the j-th component of sample vector $\mathbf{x}_i$, and the denominator N indicates the total number of samples in the given dataset.

Considering now all samples belonging to a specific class k, so with label y_k, we can define the class centroid $\mathbf{C}^k$:

$$C_j^k = \frac{\sum_{\mathbf{x}_i \in \mathcal{X} \wedge y_i = y_k} x_{ij}}{N_k} \tag{6.2}$$

where C_j^k is the j-th component of vector $\mathbf{C}^k$, x_{ij} is the j-th component of sample vector $\mathbf{x}_i$ belonging to class k ($y_i = y_k$), and the denominator N_k indicates the total number of samples in class k.

Turning the attention to a transfer case with a source and a target samples domains, $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$, respectively, we can define the "global" shift function f_s and the "per class" shift function f_s^k for a source sample $\mathbf{x}_i^s \in$ class k as follows:

$$\tilde{\mathbf{x}}_i^s = f_s(\mathbf{x}_i^s) = \mathbf{x}_i^s - \mathbf{C}_S + \mathbf{C}_T \tag{6.3}$$

$$\hat{\mathbf{x}}_i^s = f_s^k(\mathbf{x}_i^s) = \mathbf{x}_i^s - \mathbf{C}_S^k + \mathbf{C}_T^k \tag{6.4}$$

where $\tilde{\mathbf{x}}_i^s$ is the shifted source sample with "global" alignment, $\mathbf{C}_S$ and $\mathbf{C}_T$ are the total centroids of the source and target samples, respectively, whereas $\hat{\mathbf{x}}_i^s$ is the shifted source sample with "per class" alignment, $\mathbf{C}_S^k$ and $\mathbf{C}_T^k$ are the class k centroids for source and

Rosa Sicilia

target sets, respectively.

As an example, Fig. 6.6 shows the "global" alignment process in a 2-dimensional feature space, where different class clusters are distinguished by samples shape and colour, and the source and target domain are circled with a dotted line.

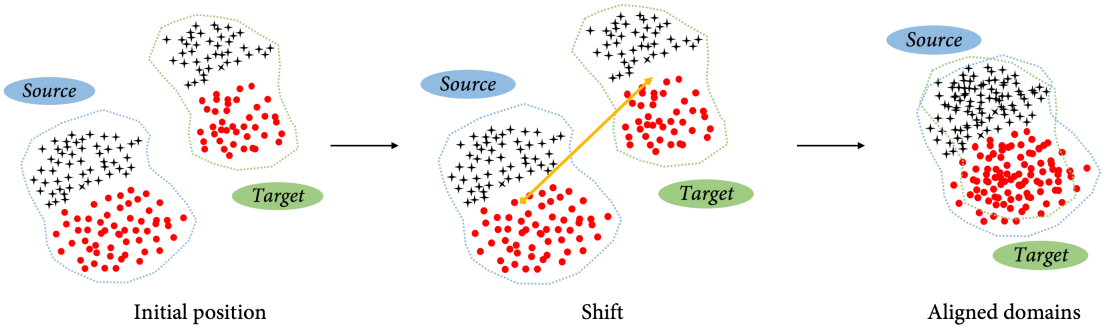


Fig. 6.6 Global domain alignment in a 2-dimensional feature space. Different class clusters are distinguished by samples shape and colour, and the source and target domain are circled with a dotted line.

6.3.2 Instance re-weighting

The samples rigid alignment just presented doesn't guarantee that the new dataset would display clear and reliable samples configurations in the feature space. Indeed the difference in source and target distributions could still affect a model performance. Therefore, our proposal presents a further step of transferring, employing instance knowledge. We used the Rule-based Characterization algorithm (RSC), presented in section 5.2.2, to analyse the aligned feature space of the training set. Indeed, this method allows to distinguish samples that belong to 8 possible arrangements in the space, characterized by different degrees of complexity for learning a classification model. Hence, we re-weighted each instance according to the category decided by the RSC tree with the weight vector $\mathbf{w} = \{w_j\}_{j=1}^{N}$, being N total number of samples in training set. The reason lies in observing that the more a configuration is complex and unreliable[2] for training, the lower is the weight w_j for an instance in that configuration.

6.3.3 System Configurations

The transfer solution presented so far is designed to be adapted to different scenarios that can occur in transfer learning problems, such as *semi-supervised* or *unsupervised* tasks. In particular, the distinction among these two cases is based on the presence or absence of local knowledge of the target domain: in other words, if the target domain

[2]The definition of complex and unreliable configuration is reported in section 5.2.2.

Rosa Sicilia

is unlabeled and so there is no local knowledge, the transfer task can be called unsupervised, whereas if there is few local knowledge, i.e. a small number of labeled target samples, the task can be considered semi-supervised.

Clearly, the above considerations are implicitly founded on the assumption that the source domain is completely labeled. However, it is worth recalling that some work in the literature define as unsupervised transfer learning the scenario where both source and target domains are unlabeled (Pan and Yang 2009). Besides this, since the analysis presented so far is based on a supervised approach for rumour detection, in the followings we will stick to the definitions reported above, considering the two possibilities with a labeled source domain.

	Task	**Transfer type**	**Alignment**	**Re-weighting**
Path 1	unsupervised	hybrid	global	only source
Path 2	semi-supervised	hybrid	per class	only source
Path 3	semi-supervised	hybrid	per class	both
Path 4	semi-supervised	hybrid	global	only source
Path 5	semi-supervised	hybrid	global	both
Path 6	unsupervised	feature-based	global	-
Path 7	unsupervised	instance-based	-	only source
Path 8	semi-supervised	feature-based	per class	-
Path 9	semi-supervised	feature-based	global	-
Path 10	semi-supervised	instance-based	-	only source
Path 11	semi-supervised	instance-based	-	both

Table 6.1 Possible configurations of the proposed approach to face different transfer scenarios. "Hybrid" refers to the use of both alignment and re-weighting steps for transfer, whereas "feature-based" and "instance-based" indicate the application only of the alignment step or the re-weighting step, respectively. The cell contents "global" and "per class" distinguish the alignment type as specified in section 6.3.1, whilst "only source" and "both" indicate the domain samples that are included in the re-weighting stage.

Table 6.1 reports eleven possible configurations of the proposed method (named as "paths"), which are distinguished by the following four characteristics:

(i) Task: it identifies the possible scenario, which could be unsupervised or semi-supervised.

(ii) Transfer type: it refers to the the type of knowledge transferred, i.e. the use of one of the blocks presented in Fig.6.5. When we use only the training alignment we apply a feature-based transfer, whereas the transfer is instance-based when using only the re-weighting block, when both blocks are used, the chain results in a hybrid transfer.

Rosa Sicilia

(iii) Alignment: this is related to the two types of data alignment presented in section 6.3.1. The string "global" indicates that we align the centroid of the whole source dataset to the centroid of the whole target dataset, whereas "per class" indicates that the alignment is performed shifting the centroid of a specific source class samples towards the centroid of the corresponding class in the target dataset.

(iv) Re-weighting: it refers to how the re-weighting step is applied to instances, e.g. if the sample set includes only source samples or both source and target samples. Straightforwardly, the presence or absence of the target samples in the set could change the configuration of instances in the feature space. This means that also the output of the RSC tree would change accordingly, so impacting on the re-weighting strategy of the set and on the knowledge transfer.

The first five paths use a hybrid transfer of knowledge, whereas the remaining five show a hyphen '-' for the unused transfer approach.

6.4 Competitors

As presented so far the proposed approach can be considered a hybrid homogeneous transfer learning method: indeed, it can combine up to a feature-based transfer step, i.e. the domain alignment, and an instance-based transfer, i.e. the RSC-based samples re-weighting. In order to analyse the efficacy of this approach we selected one competitor that belongs to the instance-based homogeneous approaches, i.e. the Transfer AdaBoost (TrAdaBoost), two methods that belong to the asymmetric feature-based homogeneous approaches, i.e. the Adaptation Regularization-based Transfer Learning (ARTL) and the Transfer Kernel Learning (TKL), and, finally, the Graph co-regularized Transfer Learning (GTL) approach that we deem falls into the symmetric feature-based homogeneous transfer learning category. This pool of methods offers a wide comparison for our approach since it includes at least one method belonging to each of the "what-to-transfer" categories used in our configurations (i.e. instance-based and feature-based) and at least one method tackling an unsupervised task and a semi-supervised task. Let's explore briefly what these methods do.

Transfer AdaBoost. As mentioned, TrAdaBoost (Dai et al. 2007) is a homogeneous transfer learning approach that transfers knowledge through instances. In particular it designed for a semi-supervised scenario where there is a small percentage of labeled target samples. This method is based on the AdaBoost classifier, which aims at boosting a weak learner accuracy by carefully adjusting the weights of the training instances (Freund and Schapire 1997). However, this method assumes that the distribution of training and test data are identical, an hypothesis that doesn't hold for transfer learning. On the

Rosa Sicilia

contrary, the Transfer AdaBoost considers both source samples and labeled target samples in the training set and specifically decreases the weight of those source samples that are wrongly predicted during training phase, weakening their impact. Indeed, the training process is the same of the AdaBoost except for the weight update at each iteration, that for source instances is multiplied by the factor $\beta^{|h_t(\mathbf{x}_i)-c(\mathbf{x}_i)|} \in (0, 1]$, where $h_t(\mathbf{x}_i)$ is the prediction for source instance $\mathbf{x}_i$, $c(\mathbf{x}_i)$ is the ground truth and $\beta \in (0, 1]$ is a constant that depends on the number of source samples and training iterations. The higher the difference in prediction $|h_t(\mathbf{x}_i) - c(\mathbf{x}_i)|$, the lower is the weight for source instance $\mathbf{x}_i$. The error in predicting such instances is a consequence of the difference in distributions between the two domains, so the re-weighting strategy allows to correct such aspect to give higher relevance to those source samples that are similar to target ones.

Adaptation Regularization-based Transfer Learning. This asymmetric feature-based transfer learning method was proposed by Long, Wang, Ding, Pan and Philip (2013) for unsupervised scenarios, i.e. with labeled source data and unlabeled target data. ARTL aims at contemporaneously correcting the difference in marginal and conditional distributions between source and target domains and at enhancing classification performance optimizing the hyperplane of a Support Vector Machine (SVM) learner. On the one hand, the classification performance enhancement is due to a manifold regularization process that optimally shifts the the SVM hyperplane (Belkin et al. 2006). On the other hand, for the distributions corrections, this framework learns an adaptive classifier simultaneously performing structural risk minimization (Vapnik 1992), reducing the marginal distribution differences with Maximum Mean Discrepancy minimization (Quanz and Huan 2009) and adjusting the conditional distributions by minimizing the differences between the source conditional probability and the target conditional probability. Since the scenario is unsupervised, the conditional distribution on the target is estimated using some pseudo-labels, which are computed using a supervised classifier learned only on the labeled source data and tested on the unlabeled target.

Transfer Kernel Learning. TKL is a kernel based approach that can be regarded as a homogeneous asymmetric feature-based transfer learning method. It is designed for an unsupervised scenario. TKL aims at learning a domain invariant kernel by directly matching the source and target distribution in the Reproducing Kernel Hilbert Space (RKHS) (Long et al. 2014). In particular, it exploits Mercer's theorem (Scholkopf and Smola 2001) to design a family of spectral kernels by extrapolating the target eigensystem on the source samples. We deem that this extrapolation produces an asymmetric transfer from the target kernel to the source kernel that reduces the marginal distribution difference in domains. The spectral kernel that minimizes the approximation error to the ground truth kernel is used to build a SVM model.

Rosa Sicilia

Graph co-regularized Transfer Learning. GTL is an homogeneous transfer learning method that is designed for unsupervised scenarios. In particular, it is based on the assumption that the input domains may share certain knowledge structures, which can be encoded in common latent factors and extracted by preserving important properties, such as statistical or geometrical (Long, Wang, Ding, Shen and Yang 2013). This approach adopts a regularized matrix factorization technique to extract common latent factors shared by source and target domains, preserving also the statistical properties of the data across domains. This common factors inference is the step that we regard as symmetric feature knowledge transfer among domains. Simultaneously to the factorization, GTL uses also a graph co-regularization technique that preserves the geometrical properties of the domains.

To sum up we considered both unsupervised and semi-supervised approaches that use instance-based or feature-based transfer learning to have a broad comparison overview for our proposal, which, indeed, can be configured to tackle diverse scenarios, as explained in the previous section 6.3.3.

6.5 Experimental Design

As first step to design the experimental setups, we identified the following five main questions:

Q.1. In the specific test case is it demanded a local knowledge on the target domain?
In other words, we would like to analyse which transfer approach could be more effective for our problem, whether the unsupervised or the semi-supervised. Answering this question could provide an important information about the labelling efforts needed for rumour detection in a new topic domain. Furthermore, this investigation essentially differs from the analysis presented in section 6.1, where we aimed at finding how much local knowledge is necessary to classify a new topic under the assumption that source and target have the same distribution. On the contrary, we now investigate the importance of local knowledge in a transfer learning context, that assumes that the source and target domains differ in their distributions, as explained in section 6.2.

Q.2. How does the proposed method perform with respect to other approaches?
This question is related to the performance of the proposed method against the competitors and against a baseline case, which doesn't apply any transfer strategy to our datasets. It is clear that resolving this question is fundamental to prove the validity of our proposal.

Q.3. Is there any difference when using a specific dataset as source or target?
We deem it would be interesting to know if there is any symmetry between the two opposite scenarios where one dataset is designed as source and the other as target, or vice

Rosa Sicilia

versa. This could eventually indicate which are possible characteristics of an effective source domain in the specific task examined.

Q.4. What type of knowledge is best transferable on our dataset?

In this case the idea is to analyse whether there is a difference among feature-based or instance-based approaches. We concentrate specifically on the differences among the hybrid, feature-based and instance-based configurations of our method, to gain a deeper understanding of its potential.

Q.5. Which is the best model that can be included in our method?

In section 6.3 we presented our transfer learning method that includes a final general classification model that has to be trained on the aligned and re-weighted instance set. We deem that it is of paramount importance to analyse different learning paradigms to identify the best suited for our task.

It is worth observing that the first three questions are general analyses that focus on the type of scenario, on the dataset and on the comprehensive performance of our proposed method. On the contrary, the last two are specific for our proposal and are introduced to finely explore its configurations.

In the following section we will describe the experimental setups designed to address these issues and that will be fully compared and discussed in section 6.6.

6.5.1 Experimental setup

To address the previous questions we designed two experiments, namely "Experiment 1" and "Experiment 2", which tackle the unsupervised and the semi-supervised tasks, respectively. Each experiment compares different configurations of baselines, competitors and proposed approach, which are necessary in the evaluation step to fully answer the five questions.

Experiment 1

The first experiment is designed to compare different setups on the unsupervised transfer learning task. In this case the only domain that is labeled and therefore can be used as training set is the source dataset. In the following we describe the configurations used as baseline, competitor and proposed approach.

Baseline 1 (B.1) This experiment is a baseline for all the unsupervised scenarios: we learn a classification algorithm on the full source dataset and evaluate it on the full target dataset. It simulates the situation of a rising rumour topic domain that hasn't been labeled yet.

Competitor 2 (C.2) In this experimental setup we run the ARTL algorithm considering the whole source dataset as training and the whole target dataset as test.

Rosa Sicilia

Competitor 3 (C.3) In this experiment we run the TKL algorithm considering the whole source dataset as training and the whole target dataset as test.

Competitor 4 (C.4) In this case we apply the GTL algorithm in the same fashion as C.2 and C.3, so with the source as training and the target as test.

Path 1 (P.1) This experiment applies the configuration named as "Path 1" presented in table 6.1. The source centroid is aligned to the target centroid regardless of the specific sample classes. Then, the source instances are re-weighted and used to train a model, that is tested on the whole target dataset.

Experiment 2

This second experiment is designed to compare different setups on the semi-supervised transfer learning task. In this case we can consider labeled also a small portion of the target dataset, therefore we define a modified 10-fold cross validation to simulate this scenario. At each iteration one target fold is used in the training set together with the entire source, whereas the remaining target folds are used as test set. We hereby describe the configurations used as baseline, competitor and proposed approach for this experiment.

Baseline 2 (B.2) This experimental setup is a baseline for the semi-supervised scenarios and it doesn't include any transfer action. Indeed, we use only the target dataset in a modified 10-fold cross validation: at each iteration one target fold is used to learn a classification model and the remaining folds are used as test. This represents a situation of a new rumour topic that has been partially labeled, so the supervised training set accounts only few samples.

Baseline 3 (B.3) This baseline is a semi-supervised experiment that includes a trivial attempt to enhance performance on the target domain by using also the source set. We perform again a modified 10-fold cross validation on the target dataset (as in B.2), but including also the whole source dataset in the training set.

Competitor 1 (C.1) This experiment refers to the application of the TrAdaBoost approach. Since it is designed for semi-supervised scenarios, we run it in a modified 10-fold cross validation fashion on the target set. As in B.3, at each iteration, the training set includes the full source and one target fold, while the test set comprises the 9 remaining target folds.

Path 3 (P.3) It refers to the configuration named as "Path 3" presented in table 6.1. In this semi-supervised case the experiment is performed in the modified 10-fold cross validation mentioned in previous experiments. At each iteration the alignment is conducted shifting all source samples towards the training target fold in a "per class" fashion, i.e. according to the specific classes. Then the training samples, that account both source and one target fold samples, are re-weighted and are used to learn the classification

Rosa Sicilia

model. The system is finally evaluated on the 9 target folds.

Path 8 (P.8) This is the configuration named as "Path 8" presented in table 6.1. As semi-supervised case the experiment is run with the modified 10-fold cross validation on target samples. At each iteration we only perform the "per class" alignment of source samples on the target training folder. Then a model is learned on the shifted source and target training and tested on the 9 target folds.

Path 11 (P.11) The last experiment refers to the "Path 11" presented in table 6.1. This is a semi-supervised approach that exploits only the re-weighting step for transferring knowledge. We use the modified 10-fold cross validation, at each iteration we re-weight the source samples together with the target training fold and train a classification model evaluated on the remaining 9 target folds.

It is worth pointing out that we didn't report here all 11 configurations of our proposal that were listed in table 6.1. After a preliminary analysis, for the sake of brevity we chose Paths 1, 3 ,8 and 11 as the most representative cases.

All the experiments were evaluated on the datasets described in chapter 3, which were alternately used both as source and target. Note that in the following experiments we consider only the binary classification task between rumour and non-rumour samples, therefore discarding from the datasets the tweets labelled as "unknown" because of their ambiguous definition, that makes these samples noisy in the transfer environment. Hence, now the #Vaccine and #Zikavirus datasets accounted a total of 990 and 694 samples, respectively.

As classification models for the baseline and the paths, we considered models belonging to different learning paradigms, including a k-Nearest Neighbour (kNN) as an instance-based classifier, a Support Vector Machine (SVM) as a kernel machine, a Decision Tree (DT) as binary tree, a Random Forest (RF) as a classification ensemble, a Multi-Layer Perceptron (MLP) as a neural network and a Linear Discriminant Analysis (LDA) as a linear classifier. As already mentioned in previous chapters, we prefer to set all the parameters to the default values, even if we acknowledge that their tuning could lead to better results. Nevertheless, the No Free Lunch Theorems for Optimization tell us that all configurations perform equally well when averaged over all possible experiments (Wolpert and Macready 1997).

The proposed method inherits a total of three parameters from the RSC algorithm, i.e. β, ε and θ for space characterization. As in the previous analysis, after preliminary experiments, we considered $\theta = 0.02$, which is a reasonable value, and also in this case its variation does not affect the results. For this reason, as in the feature selection experiments (chapter 5), we tuned only β in $(0.6, 1.0)$ and ε in $(0, 0.5)$, with a 0.1 step in both cases. These parameters are searched to determine which configuration attains the largest classification performance by using nested cross validations on all 10

Rosa Sicilia

configurations explored in table 6.1.

Besides the specific RSC parameters we needed to define the re-weight vector $\mathbf{w}$ used for the instance-based knowledge transfer (sec. 6.3.2). Recalling the observations reported in chapter 5, where we considered that the worst case in space characterization is related to overlapped regions, represented by configuration B in Fig.5.4, we decided to set to 0 weight for case B samples, whereas we assigned 1 to all others. Furthermore, since we chose a quite extreme configuration, we also introduced a parameter $f_r \in (0.0, 1.0)$, with a 0.1 step, that represents the fraction of training samples that go through the re-weighting stage. This parameter was searched together with β and ε, with the same criteria. This choice ensures that those instances that induce the presence of overlapped regions in the feature space are not influencing the transfer of knowledge, in other words that are not considered in the final training set.

Finally, for what concerns the competitors, we decided to maintain the default parameters. Indeed, as mentioned for the classification models, we preferred to maintain a baseline configuration as the basis for comparison between them. Indeed, it is reasonable that the classifier which wins on average on all the experiments would also win if a better setting was performed (Fernández et al. 2013).

6.6 Experimental results

We, now, present the results obtained running the two experiments described in the previous section: table 6.4 reports the performance of baseline configurations, table 6.5 shows the competitors' results and table 6.6 reports the proposed methods' performance. Each table shows four performance metrics for comparison: the overall accuracy, shortened as "Acc", F1 score averaged among the two classes ("F1"), recall and precision for rumour class, shortened as "Rec R" and "Prec R", respectively. As aforementioned, each experiment was executed twice: first, considering the #Vaccine dataset as source and the #Zikavirus as target, reported in the leftmost panel of each table; second, on the reversed scenario, with the #Zikavirus as source and the #Vaccine as target, shown in the rightmost panel of the tables. Moreover, for an easier comparison, we emphasised in bold those rows that report the highest performance scores in each block of experiments.

For the sake of brevity and clarity, we will now try to answer one by one the five research questions that directed our efforts, that were detailed in section 6.5.

Q.1. In the specific test case is it demanded a local knowledge on the target domain?
Let's examine the results of unsupervised and semi-supervised approaches in each table. Comparing the bold rows in table 6.4, it is possible to notice that the unsupervised configuration, i.e. B.1, shows lower performances with respect to the other two cases

Rosa Sicilia

(B.2 and B.3). This phenomenon is even more evident looking at the test case that exploits #Zikavirus dataset as source domain (left panel). Focusing on table 6.5 and table 6.6 it is possible to notice a similar behaviour: indeed, the semi-supervised approach C.1 clearly outperforms other competitors, whereas the unsupervised path P.1 shows the lowest performance among our proposed method configurations.
Hence, this recurring pattern definitely indicates that including local knowledge on the target topic helps the transfer performance. This conclusion is also consistent with the findings in section 6.1.

Q.2. How does the proposed method perform with respect to other approaches?
As an overall consideration referring to the bold rows in tables 6.4, 6.5 and 6.6, on average the proposed method shows higher performance with respect to both the baselines and the competitors. In fact, on the one hand, considering the unsupervised case, the configuration P.1 outperforms the unsupervised competitors C.2, C.3 and C.4. On the other hand, with the semi-supervised scenarios P.3, P.8 and P.11, the proposed method definitely outperforms the competitor TrAdaBoost (C.1), however it reaches equal or slightly higher performance confronting it to the baselines B.2 and B.3.
It is worth noticing that the all the competitors show lower performance also with respect to the baselines. This can be due to a phenomenon usually named as *negative transfer*, which happens when the application of a transfer learning strategy decreases the performance the target dataset rather than increasing them. For what concerns the proposed method, instead, since the final results are almost equal to the baselines, a reader can argue that we are witnessing a lack of transfer and that the method is not useful. To prove that this is not the case, we increase the local knowledge on the target domain running configuration P.3, B.2 and B.3 of Experiment 2 in a modified 5-fold cross validation. For brevity we only report in Table 6.2 the best accuracy using the Random Forest classifier. As it is possible to notice increasing the local knowledge on the target led to an improvement in all configurations and now P.3 outperforms B.2 in one test case and B.3 in both cases. This means that a slight transfer exists. We deem that a finer tuning of the parameter could further improve this results, as we also report in the conclusions, since this can be a direction for further investigation. Note also that the performance achieved in Table 6.2 on the #Vaccine dataset (right panel) outperform an average human annotator (Table 4.6); furthermore, comparing this result with the corresponding one reported in section 6.1.2 we notice that in this case we need less samples from the target dataset to outperform the human, suggesting that transfer learning could be beneficial.

Beyond this analysis, we further examine the negative transfer results of the competitors

Rosa Sicilia

	#Vaccine (S) - #Zikavirus (T)	#Zikavirus (T) - #Vaccine (S)
P.3	0.87	0.84
B.2	0.87	0.83
B.3	0.86	0.83

Table 6.2 Accuracy of configurations P.3, B.2 and B.3 run in a modified 5-fold cross validation.

C.2, C.3 and C.4 on our datasets. In this respect to provide a fairer comparison, we considered the test data from Reuters-21578[3] text database used for evaluation in all three reference papers where the competitors were presented (Long et al. 2014, Long, Wang, Ding, Shen and Yang 2013, Long, Wang, Ding, Pan and Philip 2013). Table 6.3 reports the accuracy on three combinations of source/target pair of Reuters text data: org vs people, org vs place and people vs place. Each value in the table is computed as the mean performance among each pair and the inverted one, e.g. for the first case: the reported value is the average between the accuracy on org vs people and the accuracy on people vs org. Since in the references of the competitors there weren't accuracy tables for all datasets, we replicated the experiments using their source codes with the same parameter used in the papers[4]. The performance achieved by the proposed method was obtained running the selected configurations with the setups used to compute the results in table 6.6. We hereby report only the performance achieved considering the Random Forest classifier, as best case results. As it is possible to see, the proposed method gains comparable performance to the competitors, even though its parameters are not finely tuned for the specific task. This finding shows that the low performance of the competitors is related to the complexity of the specific transfer task, rather than to the fine tuning of the algorithms.

	org vs people	org vs place	people vs place
P.1	0.77	0.66	0.58
P.3	0.81	0.77	0.71
P.8	0.82	0.77	0.70
P.11	0.80	0.73	0.62
C.2	0.82	0.70	0.57
C.3	0.84	0.81	0.68
C.4	0.84	0.76	0.66

Table 6.3 Proposed method and competitor accuracy on three baseline datasets of the Reuters-21578 database used in Long et al. (2014), Long, Wang, Ding, Shen and Yang (2013), Long, Wang, Ding, Pan and Philip (2013) for competitors evaluation.

[3]http://www.daviddlewis.com/resources/testcollections/reuters21578/.
[4]The source code and datasets are publicly available at http://ise.thss.tsinghua.edu.cn/ mlong/

Rosa Sicilia

Q.3. Is there any difference when using a specific dataset as source or target?
Comparing the right and left panels of all tables (tab. 6.4, 6.5 and 6.6) it is possible to notice an overall asymmetric behaviour between the two test cases: indeed, the using the #Vaccine dataset as source seems to provide better performance than using the #Zikavirus. We deem that this phenomenon is due to two factors: first, the sample sizes, indeed the #Vaccine dataset accounts more than the 30% of #Zikavirus data; second, the labels stability given by the gold standard, which was computed only on the #Vaccine dataset.

Q.4. What type of knowledge is best transferable on our dataset? In all our experiments we considered two type of knowledge to transfer: intance-based and feature-based. Focusing on the proposed method (Table 6.6), there is no real difference among the hybrid case (P.3), the feature-based (P.8) and the instance-based (P.11). In addition, all three cases outperform the competitors.
On these grounds, we could conclude that both approaches, instance and feature-based, are individually robust. However, it is worth noticing that in competitors table 6.5, the only instance-based approach (C.1) clearly wins on the others. This finding indicates that probably the instance-based knowledge it's preferable to transfer in the specific test case.

Q.5. Which is the best model that can be included in our method?
Examining table 6.4 and table 6.6, the learning paradigm that shows a clear win is the Random Forest (RF). We deem that this is probably related to two peculiarities of this classifier: (i) its embedded capability to perform feature selection, a step that has been left out from these experiments; (ii) the fact that it is the only ensemble learner among the chosen methods.

6.7 Take-home messages

In this chapter we tackled the problem of creating a rumour detection system effective on different and new topics. Indeed, once a system has been trained on a specific dataset, it isn't trivial to discover rumours in a novel rising topic. For this reason we conducted a preliminary investigation in order to test the cross-topic robustness of the system developed so far. This led us to the conclusion that given two different topics, a training topic and a test topic, the model needs at least the 80% of the test topic in the training set to provide acceptable performance in classifying the remaining test samples. Hence, considered the huge effort of the labeling process, this is not an acceptable solution in a real life rumour detection environment.

Rosa Sicilia

Baselines		#Vaccine (S) - #Zikavirus (T)				#Zikavirus (S) - #Vaccine (T)			
		Acc	F1	Rec R	Prec R	Acc	F1	Rec R	Prec R
B.1	kNN	0.36	0.26	0.00	0.00	0.40	0.29	1.00	0.40
	SVM (rbf)	0.36	0.26	0.00	0.00	0.40	0.29	1.00	0.40
	SVM (linear)	0.36	0.26	0.00	0.00	0.39	0.29	0.98	0.39
	DT	**0.82**	**0.81**	**0.82**	**0.90**	**0.45**	**0.44**	**0.38**	**0.33**
	RF	0.64	0.64	0.50	0.88	0.39	0.38	0.67	0.36
	MLP	0.55	0.52	0.61	0.66	0.41	0.38	0.76	0.38
	LDA	0.53	0.51	0.57	0.65	0.49	0.47	0.82	0.43
B.2	kNN	0.82	0.81	0.83	0.88	0.63	0.62	0.54	0.54
	SVM (rbf)	0.83	0.81	0.85	0.88	0.60	0.38	0.00	0.80
	SVM (linear)	0.82	0.81	0.80	0.91	0.47	0.39	0.57	0.35
	DT	0.79	0.77	0.83	0.85	0.69	0.68	0.60	0.62
	RF	**0.84**	**0.82**	**0.89**	**0.87**	**0.77**	**0.75**	**0.62**	**0.75**
	MLP	0.77	0.75	0.81	0.83	0.62	0.60	0.53	0.53
	LDA	0.77	0.76	0.76	0.87	0.64	0.62	0.48	0.56
B.3	kNN	0.83	0.82	0.84	0.89	0.62	0.60	0.52	0.53
	SVM (rbf)	0.81	0.81	0.78	0.92	0.61	0.39	0.01	0.80
	SVM (linear)	0.64	0.39	1.00	0.64	0.54	0.53	0.59	0.46
	DT	0.82	0.81	0.86	0.86	0.69	0.68	0.65	0.61
	RF	**0.84**	**0.83**	**0.89**	**0.87**	**0.76**	**0.74**	**0.65**	**0.73**
	MLP	0.78	0.77	0.76	0.88	0.64	0.62	0.56	0.54
	LDA	0.64	0.63	0.62	0.78	0.55	0.53	0.46	0.43

Table 6.4 Performance of the baselines experiments described in section 6.5. The right panel reports the results with the #Zikavirus dataset as source domain (S) and the #Vaccine as target (T), whereas the left panel shows what happens with the inverted roles.

Competitors	#Vaccine (S) - #Zikavirus (T)				#Zikavirus (S) - #Vaccine (T)			
	Acc	F1	Rec R	Prec R	Acc	F1	Rec R	Prec R
C.1	**0.74**	**0.74**	**0.61**	**0.98**	**0.61**	**0.43**	**0.07**	**0.54**
C.2	0.36	0.26	0.00	0.00	0.40	0.29	1.00	0.40
C.3	0.36	0.26	0.00	0.00	0.40	0.29	1.00	0.40
C.4	0.42	0.35	0.25	0.52	0.48	0.38	0.76	0.39

Table 6.5 Performance of the competitors chosen in section 6.4 and the experimental setups described in section 6.5. The right panel reports the results with the #Zikavirus dataset as source domain (S) and the #Vaccine as target (T), whereas the left panel shows what happens with the inverted roles.

Rosa Sicilia

		#Vaccine (S) - #Zikavirus (T)				#Zikavirus (S) - #Vaccine (T)			
		Acc	F1	Rec R	Prec R	Acc	F1	Rec R	Prec R
P.1	kNN	0.36	0.26	0.00	0.00	0.39	0.29	0.97	0.40
	SVM (rbf)	0.36	0.26	0.00	0.00	0.40	0.29	1.00	0.40
	SVM (linear)	0.36	0.26	0.00	0.00	**0.62**	**0.60**	**0.56**	**0.52**
	DT	0.44	0.38	0.57	0.56	0.43	0.42	0.33	0.31
	RF	**0.62**	**0.40**	**0.96**	**0.63**	0.40	0.40	0.55	0.34
	MLP	0.49	0.48	0.30	0.74	0.39	0.31	0.94	0.39
	LDA	0.61	0.48	0.85	0.64	0.60	0.43	0.06	0.53
P.3	kNN	0.82	0.81	0.85	0.87	0.63	0.62	0.55	0.54
	SVM (rbf)	0.77	0.75	0.72	0.82	0.61	0.39	0.01	1.00
	SVM (linear)	0.64	0.39	1.00	0.64	0.51	0.48	0.61	0.45
	DT	0.82	0.80	0.87	0.85	0.69	0.68	0.64	0.62
	RF	**0.84**	**0.83**	**0.90**	**0.87**	**0.77**	**0.75**	**0.61**	**0.77**
	MLP	0.77	0.75	0.78	0.85	0.64	0.62	0.52	0.55
	LDA	0.66	0.66	0.57	0.86	0.57	0.55	0.46	0.47
P.8	kNN	0.82	0.80	0.83	0.88	0.64	0.62	0.54	0.55
	SVM (rbf)	0.82	0.81	0.79	0.91	0.61	0.39	0.01	0.90
	SVM (linear)	0.64	0.39	1.00	0.64	0.51	0.49	0.54	0.44
	DT	0.79	0.77	0.84	0.84	0.69	0.68	0.62	0.62
	RF	**0.85**	**0.84**	**0.89**	**0.88**	**0.78**	**0.76**	**0.64**	**0.78**
	MLP	0.79	0.78	0.81	0.86	0.64	0.63	0.58	0.55
	LDA	0.66	0.65	0.56	0.85	0.57	0.55	0.46	0.46
P.11	kNN	0.82	0.80	0.82	0.89	0.62	0.60	0.51	0.53
	SVM (rbf)	0.81	0.80	0.78	0.91	0.61	0.39	0.01	1.00
	SVM (linear)	0.53	0.34	0.60	0.38	0.54	0.50	0.47	0.43
	DT	0.81	0.80	0.86	0.86	0.69	0.68	0.66	0.61
	RF	**0.84**	**0.82**	**0.87**	**0.88**	**0.76**	**0.74**	**0.62**	**0.74**
	MLP	0.78	0.77	0.79	0.86	0.64	0.62	0.54	0.55
	LDA	0.64	0.63	0.63	0.78	0.56	0.54	0.46	0.45

Table 6.6 Performance of 4 of the proposed method configurations described in section 6.5. The right panel reports the results with the #Zikavirus dataset as source domain (S) and the #Vaccine as target (T), whereas the left panel shows what happens with the inverted roles.

Rosa Sicilia

On these grounds, we decided to direct our focus towards transfer learning techniques, proposing a novel hybrid solution that as first step aligns the two topic domains in the feature space (feature-based transfer) and then exploits the RSC algorithm to reweight the training instances according to their configuration (instance-based transfer). We studied several setups of our method and tested them also on our two health related datasets. The results were compared with four competitors and three baselines, showing promising performance particularly when our method was combined with the Random Forest classifier. An interesting finding was the negative transfer phenomenon shown by the competitors on our datasets. To provide a fairer comparison we also evaluated the proposed solution on three benchmark text dataset used by the competitors to validate their methods. This experiment proved the efficacy of our proposal also on a different pool of data.

Furthermore we analysed our hybrid solution compared with the configurations exploiting only the feature-based or the instance-based transfers. This didn't reveal a particular difference in the performance, meaning that the single knowledge transfers are equally robust in the specific test cases.

The experiments confirmed that having in the training set few local knowledge of the test topic, that in transfer learning is referred to as target domain, is particularly relevant for performance enhancement. Moreover, even though these results aren't directly comparable with those of the preliminary analysis, we were able to achieve over the 80% of accuracy using only the 10% of the target domain samples and to outperform the human average performance with only the 20% of target samples. Increasing target local knowledge to the 20% yield also to outperform baselines experiment, showing the transfer efficacy of the proposed solution. As a direction for further investigation, we deem that it would be interesting to perform a finer tuning of the parameter of our model, since it could further improve these results.

Finally, another important finding was that the #Vaccine dataset performs particularly well as source dataset, as also seen in the preliminary analysis when used as training topic. This indicates that the sample size and the labeling process have a strong impact in the transferring process.

Rosa Sicilia

CHAPTER 7

"..there's still a long way to go"

In this work we tackled the problem of rumour detection at the micro-level on Twitter social microblog under different perspectives. This research field accounts many contributions focused on macro-level analyses, whereas there is few work on the detection of rumour single posts (Zubiaga et al. 2017, Fard et al. 2019). As widely discussed in chapters 1 and 2, being able to recognize rumours at the micro-level is particularly useful in many domains such as security and health, which is a field where rumours show a significant concentration (Wu et al. 2015). For this reason we deem that it could have a beneficial impact on people quality of life, if we could provide users with the reliability level of the health information they look for and share in social microblogs.

Having stressed again these concepts, in this chapter we would like to provide a summary of the contributions and findings of this work and of the future directions to strike out.

7.1 The road covered so far

Walking through the main steps of this work, as first contribution, we created two new health-related datasets for rumour detection on Twitter, i.e. the #Zikavirus dataset and the #Vaccine dataset. These were manually labelled into three categories, namely rumour, non-rumour and unknown, and thanks to the further effort of three independent annotators, we were also able to provide a gold standard for the #Vaccine dataset.

Then as a second major step, we presented and validate a novel feature set exploiting not only features available in the literature, but also new descriptors inspired by the study of the graph theory and the social influence models. These included the likelihood that a tweet is retweeted, the likelihood of a URL to be shared, the conversation size, the fraction of users followers of root, and the fraction of tweets with URLs. With this feature set we were able to achieve an accuracy of 82.3% on the #Zikavirus dataset and of 96.0% on the #Vaccine dataset, outperforming also human annotators.

The third step in this path was the development of a novel feature selection filter, relying on a new Rule-based Space Characterization algorithm (RSC), also developed in this work. This method aims at identifying samples lying in complex and less reliable

Rosa Sicilia

regions in the feature space, selecting the feature subset that reduces the number of instances in those regions. Comparing our approach with other well-established feature selection techniques on two Twitter health datasets, the results showed that in most of the combinations our proposal outperformed the competitors. Moreover, the system that exploited the novel RSC filter was able to enhance the performance achieved on both datasets, obtaining an accuracy of 87.4% on the #Zikavirus set and of 96.8% on the #Vaccine set.

Finally, as last contribution we explored the potential of transfer learning for our task, proposing an hybrid transfer method. This combined a feature-based transfer, implemented as a domain alignment step, with an instance-based transfer, that exploited the RSC algorithm to re-weight the instances. We compared our proposal with four state-of-the-art transfer learning competitors and three baseline configurations, obtaining promising results. Indeed, we were able to outperform the competitors, showing also comparable performance on three of their benchmark datasets. This proved not only the validity of the proposed solution, but also the potential and versatility of the RSC algorithm that was effective for two very different goals, i.e. feature selection and transfer learning.

7.2 Future directions

Recalling the road covered so far, probably the first next step forward could be to go beyond our health related data and to validate our systems on different datasets, with novel topics. This would let to further evaluate our transfer learning proposal with a finer tuning and study the system's transfer power. Moreover with new datasets we could also explore transfer among more than two domains, in a multi-transfer environment, or rather to understand whether the source and target domains could belong to different higher level topics or not. Besides this, with a suitable sample size, also deep learning-based transfer learning could be a reasonable path to undertake.

A different direction could be to drive further the network analyses, exploiting the micro-level approach to discover information about eventual polarizations of the social networks, i.e. the tendency of people to immerse themselves in social circles in such a way that they are primarily exposed to contents that agree with their beliefs (Kumar and Shah 2018). Indeed, this could give important information about rumour impact and diffusion, being a considerable feature in the rumour detection process.

Another step forward in rumour analyses could be directed to modeling the temporal aspect of the diffusion. This wasn't studied in this work, but it is nonetheless a fundamental factor that could improve the developed models.

Finally, beyond the rumour detection task, it would be interesting to explore the potential applications of the RSC algorithm. Broadly speaking, the proposed approach can

Rosa Sicilia

be viewed as a model for explanatory artificial intelligence (XAI), as it is able to summarize the reasons for classifier behaviour producing insights about the causes of their decisions (Gilpin et al. 2018). Hence, with further analysis such space characterization algorithm could potentially give information about the reliability of the classification or rather the suited re-sampling strategy in imbalanced learning contexts.

Rosa Sicilia

REFERENCES

Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G. (2008), Finding high-quality content in social media, *in* 'Proceedings of the 2008 International Conference on Web search and Data mining', ACM, pp. 183–194.

Alonso, O., Carson, C., Gerster, D., Ji, X. and Nabar, S. U. (2010), Detecting uninteresting content in text streams, *in* 'SIGIR Crowdsourcing for Search Evaluation Workshop'.

Belkin, M., Niyogi, P. and Sindhwani, V. (2006), 'Manifold regularization: A geometric framework for learning from labeled and unlabeled examples', *Journal of machine learning research* **7**(Nov), 2399–2434.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg.

Castillo, C., Mendoza, M. and Poblete, B. (2011), Information credibility on Twitter, *in* 'Proceedings of the 20th international conference on World Wide Web', ACM, pp. 675–684.

Chung, S. S. and Zhang, S. (2017), 'Volatility estimation using support vector machine: Applications to major foreign exchange rates', *Electronic Journal of Applied Statistical Analysis* **10**(2), 499–511.

Dai, W., Yang, Q., Xue, G.-R. and Yu, Y. (2007), Boosting for transfer learning, *in* 'Proceedings of the 24th international conference on Machine learning', ACM, pp. 193–200.

Dave, K. S., Bhatt, R. and Varma, V. (2011), Modelling Action Cascades in Social Networks., *in* 'ICWSM'.

DiFonzo, N. and Bordia, P. (2007), *Rumor psychology: social and organizational approaches.*, American Psychological Association.

Esuli, A. and Sebastiani, F. (2006), Sentiwordnet: A publicly available lexical resource for opinion mining, *in* 'Proceedings of LREC', Vol. 6, Citeseer, pp. 417–422.

Rosa Sicilia

Fard, A. E., Mohammadi, M., Chen, Y. and Van de Walle, B. (2019), 'Computational rumor detection without non-rumor: A one-class classification approach', *IEEE Transactions on Computational Social Systems* **6**(5), 830–846.

Fernández, A., López, V., Galar, M., Del Jesus, M. J. and Herrera, F. (2013), 'Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches', *Knowledge-Based Systems* **42**, 97–110.

Freeman, L. C. (1978), 'Centrality in social networks conceptual clarification', *Social networks* **1**(3), 215–239.

Freund, Y. and Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of computer and system sciences* **55**(1), 119–139.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. and Kagal, L. (2018), Explaining explanations: An overview of interpretability of machine learning, *in* '2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)', IEEE, pp. 80–89.

Guan, W., Gao, H., Yang, M., Li, Y., Ma, H., Qian, W., Cao, Z. and Yang, X. (2014), 'Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events', *Physica A: Statistical Mechanics and its Applications* **395**, 340–351.

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

He, H. and Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on knowledge and data engineering* **21**(9), 1263–1284.

Ho, Y.-C. and Pepyne, D. L. (2002), 'Simple explanation of the no-free-lunch theorem and its implications', *Journal of optimization theory and applications* **115**(3), 549–570.

Huang, S. H. (2015), 'Supervised feature selection: A tutorial.', *Artif. Intell. Research* **4**(2), 22–37.

Hughes, A. L. and Palen, L. (2009), 'Twitter adoption and use in mass convergence and emergency events', *International Journal of Emergency Management* **6**(3-4), 248–260.

Kononenko, I. (1994), Estimating attributes: analysis and extensions of relief, *in* 'European conference on machine learning', Springer, pp. 171–182.

Kumar, S. and Shah, N. (2018), 'False information on web and social media: A survey', *arXiv preprint arXiv:1804.08559* .

Rosa Sicilia

Kwon, S. and Cha, M. (2014), Modeling Bursty Temporal Pattern of Rumors., *in* 'ICWSM'.

Kwon, S., Cha, M. and Jung, K. (2017), 'Rumor detection over varying time windows', *PloS one* **12**(1), e0168344.

Kwon, S., Cha, M., Jung, K., Chen, W. and Wang, Y. (2013), Prominent features of rumor propagation in online social media, *in* 'Data Mining (ICDM), 2013 IEEE 13th International Conference on', IEEE, pp. 1103–1108.

Liu, H., Setiono, R. et al. (1996), A probabilistic approach to feature selection-a filter solution, *in* 'ICML', Vol. 96, Citeseer, pp. 319–327.

Long, M., Wang, J., Ding, G., Pan, S. J. and Philip, S. Y. (2013), 'Adaptation regularization: A general framework for transfer learning', *IEEE Transactions on Knowledge and Data Engineering* **26**(5), 1076–1089.

Long, M., Wang, J., Ding, G., Shen, D. and Yang, Q. (2013), 'Transfer learning with graph co-regularization', *IEEE Transactions on Knowledge and Data Engineering* **26**(7), 1805–1818.

Long, M., Wang, J., Sun, J. and Philip, S. Y. (2014), 'Domain invariant transfer kernel learning', *IEEE Transactions on Knowledge and Data Engineering* **27**(6), 1519–1532.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F. and Cha, M. (2016), Detecting rumors from microblogs with recurrent neural networks., *in* 'IJCAI', pp. 3818–3824.

Ma, J., Gao, W., Wei, Z., Lu, Y. and Wong, K.-F. (2015), Detect rumors using time series of social context information on microblogging websites, *in* 'Proceedings of the 24th ACM International on Conference on Information and Knowledge Management', ACM, pp. 1751–1754.

Pan, S. J. and Yang, Q. (2009), 'A survey on transfer learning', *IEEE Transactions on knowledge and data engineering* **22**(10), 1345–1359.

Perrin, A. (2015), 'Social media usage: 2005-2015'.

Qazvinian, V., Rosengren, E., Radev, D. R. and Mei, Q. (2011), Rumor has it: identifying misinformation in microblogs, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, pp. 1589–1599.

Rosa Sicilia

Quanz, B. and Huan, J. (2009), Large margin transductive transfer learning, *in* 'Proceedings of the 18th ACM conference on Information and knowledge management', ACM, pp. 1327–1336.

Saeys, Y., Inza, I. and Larrañaga, P. (2007), 'A review of feature selection techniques in bioinformatics', *bioinformatics* **23**(19), 2507–2517.

Scholkopf, B. and Smola, A. J. (2001), *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press.

Shearer, E. and Gottfried, J. (2017), 'News use across social media platforms 2017', *Pew Research Center* **7**.

Tolosi, L., Tagarev, A. and Georgiev, G. (2016), An analysis of event-agnostic features for rumour classification in twitter, *in* 'Tenth International AAAI Conference on Web and Social Media'.

Vapnik, V. (1992), Principles of risk minimization for learning theory, *in* 'Advances in neural information processing systems', pp. 831–838.

Vidyasagar, M. (2002), *A theory of learning and generalization*, Springer-Verlag New York, Inc.

Wang, Y., McKee, M., Torbica, A. and Stuckler, D. (2019), 'Systematic literature review on the spread of health-related misinformation on social media', *Social Science and Medicine* p. 112552.

Wang, Z., Guo, Y., Wang, J., Li, Z. and Tang, M. (2019), 'Rumor events detection from chinese microblogs via sentiments enhancement', *IEEE Access* **7**, 103000–103018.

Weiss, K., Khoshgoftaar, T. M. and Wang, D. (2016), 'A survey of transfer learning', *Journal of Big data* **3**(1), 9.

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Wolpert, D. H. and Macready, W. G. (1997), 'No Free Lunch Theorems for Optimization', *IEEE Transactions on Evolutionary Computation* **1**(1), 67–82.

Wu, K., Yang, S. and Zhu, K. Q. (2015), False rumors detection on Sina Weibo by propagation structures, *in* 'IEEE 31st International Conference on Data Engineering', IEEE, pp. 651–662.

Rosa Sicilia

Yang, F., Liu, Y., Yu, X. and Yang, M. (2012), Automatic detection of rumor on Sina Weibo, *in* 'Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics', ACM, p. 13.

Zeng, L., Starbird, K. and Spiro, E. S. (2016), # unconfirmed: Classifying rumor stance in crisis-related social media messages, *in* 'Tenth International AAAI Conference on Web and Social Media'.

Zhao, Z., Resnick, P. and Mei, Q. (2015), Enquiring minds: Early detection of rumors in social media from enquiry posts, *in* 'Proceedings of the 24th International Conference on World Wide Web', International World Wide Web Conferences Steering Committee, pp. 1395–1405.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. and Procter, R. (2018), 'Detection and resolution of rumours in social media: A survey', *ACM Computing Surveys (CSUR)* **51**(2), 32.

Zubiaga, A., Liakata, M. and Procter, R. (2017), Exploiting context for rumour detection in social media, *in* 'International Conference on Social Informatics', Springer, pp. 109–123.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S. and Tolmie, P. (2016), 'Analysing how people orient to and spread rumours in social media by looking at conversational threads', *PloS One* **11**(3), e0150989.

Rosa Sicilia

LIST OF PUBLICATIONS

1. Sicilia, R., Giudice, S. L., Pei, Y., Pechenikiy, M., and Soda, P. (2018). Twitter Rumour Detection in the Health Domain. *Expert Systems with Applications*.

2. Sicilia, R., Merone, M., Valenti, R., Cordelli, E., D'Antoni, F., De Ruvo, V., ... and Soda, P. (2018, December). Cross-topic Rumour Detection in the Health Domain. *In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2056-2063). IEEE.

3. R. Sicilia, S. Lo Giudice, Y. Pei, M. Pechenizkiy, P. Soda, Health-related Rumour Detection On Twitter, *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017.

Rosa Sicilia

Appendices

Rosa Sicilia

Appendix A

Feature set Definition

We hereby report a formal description of the 24 features used in our work. For each tweet (T) in the dataset we used the following symbols:

- U indicates the current user author of the post (T).
- C indicates the conversation graph containing T.
- F_i refers to the i-th follower of U.
- f_i refers to the i-th account followed by U.
- s_i is the i-th status of U.
- t_p is the time at which T was posted.
- t_0 is the time at which user U created the account.
- RT_i is the i-th retweet of the current post.
- url_i^T is the i-th occurrence in the database of the url attached to T.
- S_i is the sentiment associated to the i-th word in T.
- $Root(C)$ identifies the root user in a conversation C.

A.1 Influence Potential features

A.1.1 User level:

1. Number of followers ($\#Followers$):

$$\#Followers(U) = \sum_i F_i \tag{A.1}$$

2. Number of followings ($\#Followings$):

$$\#Followings(U) = \sum_i f_i \tag{A.2}$$

3. Number of Statuses ($\#Statuses$):

$$\#Statuses(U) = \sum_i s_i \tag{A.3}$$

Rosa Sicilia

4. Age of account registration expressed in months ($RegistrationAge$):

$$\#RegistrationAge(U) = t_p - t_0 \tag{A.4}$$

5. Is a Retweet? ($isRT$): binary value that indicates if T is actually a retweet (1) or not (0).
6. has URL? ($hasURL$): binary value that indicates if T contains at least one url attached (1) or not (0).
7. Has question mark? ($has\text{“?”}?$): binary value indicating if the current post contains at least one question mark (1) or not (0).
8. Probability of a retweet (P_{rt}): probability of a tweet to be retweeted.

$$P_{rt} = \frac{\sum_i RT_i}{\sum_i T_i} \tag{A.5}$$

9. Probability of a url (P_{url}): probability of a url to be shared.

$$P_{url} = \frac{\sum_i url_i^T}{\sum_{i,T} url_i^T} \tag{A.6}$$

10. is Follower?: binary value that indicates if U is follower of the root user in a conversation (1) or not (0).

A.1.2 Network level:

1. Average number of followers ($Avg\#Followers$):

$$Avg\#Followers(C) = \frac{\sum_{U \in C} \#Followers(U)}{size(C)} \tag{A.7}$$

2. Average number of followings ($Avg\#Followings$):

$$Avg\#Followings(C) = \frac{\sum_{U \in C} \#Followings(U)}{size(C)} \tag{A.8}$$

3. Average number of statuses ($Avg\#Statuses$):

$$Avg\#Statuses(C) = \frac{\sum_{U \in C} \#Statuses(U)}{size(C)} \tag{A.9}$$

Rosa Sicilia

4. Average registration age ($AvgRegistrationAge$):

$$AvgRegistrationAge(C) = \frac{\sum_{U \in C} RegistrationAge(U)}{size(C)} \tag{A.10}$$

A.2 Personal Interest

A.2.1 User Level:

1. Sentiment Score:

$$SentimentScore(T) = \sum_{i} S_i \tag{A.11}$$

A.2.2 Network level:

1. Average Sentiment Score (Avg Sentiment Score):

$$AvgSentimentScore(C) = \frac{\sum_{T \in C} SentimentScore(T)}{size(C)} \tag{A.12}$$

2. Fraction of positive:

$$AvgSentimentScore(C) = \frac{\sum_{T \in C} SentimentScore(T)}{size(C)} \tag{A.13}$$

   if $SentimentScore(T) > 0$.

3. Fraction of negative:

$$AvgSentimentScore(C) = \frac{\sum_{T \in C} SentimentScore(T)}{size(C)} \tag{A.14}$$

   if $SentimentScore(T) < 0$.

A.3 Network Characteristics

The formal definition of Page rank, Betweenness Centrality and Closeness Centrality is reported in Freeman (1978).

Rosa Sicilia

A.3.1 Network level:

1. Conversation size: it is the number of distinct users involved in a conversation.

$$size(C) = \sum_i U_i \tag{A.15}$$

2. Fraction of Users Followers of the Root ($FUFR$):

$$FUFR(C) = \frac{\#Followers(Root(C))_{F_i \in C}}{size(C)} \tag{A.16}$$

3. Fraction of Tweets with URL ($FTWU$):

$$FTWU(C) = \frac{\sum_{T \in C} T}{size(C)} \tag{A.17}$$

if T contains at least 1 url.

Rosa Sicilia

Appendix B

Other work

B.1 Multivariate Time Series Forecasting

Reference Soda, P., Sicilia, R., Acciai, L., and Iannello, G. (2019). Grasping inter-attribute and temporal variability in multivariate time series. IEEE Transactions on Big Data.

Abstract The rising capabilities of storing and registering data has increased the number of temporal datasets, boosting the attention on time series classification and forecasting. In case of multivariate time series, symbolic methods that try to predict phenomena transform the data into a more compact format to produce a representation of the time series easy to be handled in a machine learning framework. However, up to now these representations do not grasp information on both inter-attribute variability and temporal variability. In this work we present an approach that, taking into account the relationships between attributes and their periodicity, reduces the multivariate time series to a collection of symbols, whose distribution is represented by histograms (represented in Fig. B.1). The approach has been successfully tested on a publicly available dataset, the Telecom Italia Big Data Challenge 2014 dataset, reporting also the results attained by other methods available in the literature.

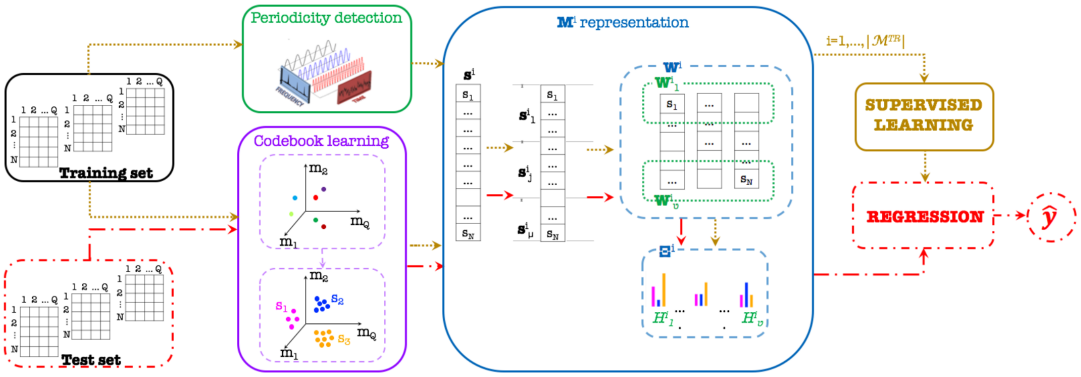


Fig. B.1 Pipeline of the proposed method.

Rosa Sicilia

B.2 Radiomics

(1) Reference Ramella, S., Fiore, M., Greco, C., Cordelli, E., Sicilia, R., Merone, M., *et al.* (2018). A radiomic approach for adaptive radiotherapy in non-small cell lung cancer patients. PloS one, 13(11), e0207455.

Abstract The primary goal of precision medicine is to minimize side effects and optimize efficacy of treatments. Recent advances in medical imaging technology allow the use of more advanced image analysis methods beyond simple measurements of tumor size or radio- tracer uptake metrics. The extraction of quantitative features from medical images to characterize tumor pathology or heterogeneity is an interesting process to investigate, in order to provide information that may be useful to guide the therapies and predict survival. This paper discusses the rationale supporting the concept of radiomics and the feasibility of its application to Non-Small Cell Lung Cancer in the field of radiation oncology research (Figure B.2 shows a region of interest example in a 3D image). We studied 91 stage III patients treated with concurrent chemoradiation and adaptive approach in case of tumor reduction during treatment. We considered 12 statistics features and 230 textural features extracted from the CT images. In our study, we used an ensemble learning method to classify patients' data into either the adaptive or non-adaptive group during chemoradiation on the basis of the starting CT simulation. Our data supports the hypothesis that a specific signature can be identified (AUC 0.82, as reported in Figure B.3). In our experience, a radiomic signature mixing semantic and image-based features has shown promising results for personalized adaptive radiotherapy in non-small cell lung cancer.

Rosa Sicilia

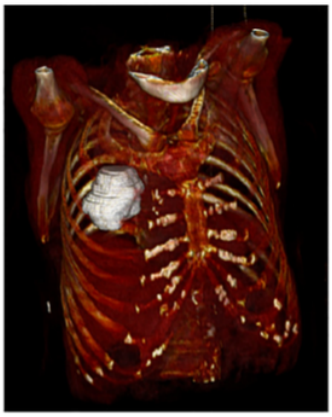


Fig. B.2 Example of region of interest in a 3D image.

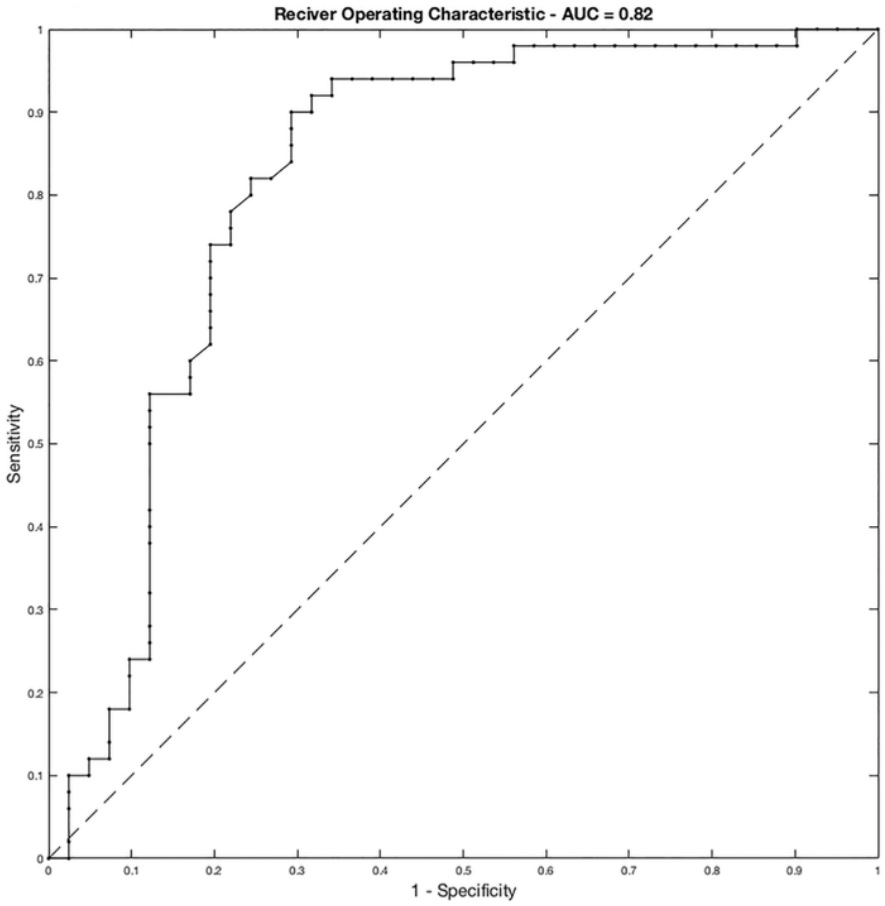


Fig. B.3 ROC curve of the proposed system.

Rosa Sicilia

(2) Reference D'Amico, N. C., Merone, M., Sicilia, R., Cordelli, E., D'Antoni, F., Zanetti, I. B., *et al.* (2019). Tackling imbalance radiomics in acoustic neuroma. International Journal of Data Mining and Bioinformatics, 22(4), 365-388.

Abstract Acoustic neuroma is a primary intracranial tumour of the myelinforming cells of the 8th cranial nerve. Although it is a slow growing benign tumour, symptoms in the advanced phase can be serious. Hence, controlling tumour growth is essential and stereotactic radiosurgery, which can be performed with the CyberKnife robotic device, has proven effective for managing this disease. However, this approach may have side effects and a follow-up is necessary to assess its efficacy. To optimise the administration of this treatment, in this work we present a machine learning-based radiomics approach that first computes quantitative biomarkers from MR images routinely collected before the CyberKnife treatment and then predicts the treatment response. To tackle the challenge of class imbalance observed in the available dataset we present a cascade of cost-sensitive decision trees. The presented pipeline is depicted in Figure B.4. We also experimentally compare the proposed approach with several approaches suited for learning under class skew. The results achieved, with a global accuracy of 0.92, demonstrate that radiomics has a great potential in predicting patients response to radiosurgery prior to the treatment that, in turns, can reflect into great advantages in therapy planning, sparing radiation toxicity and surgery when unnecessary.

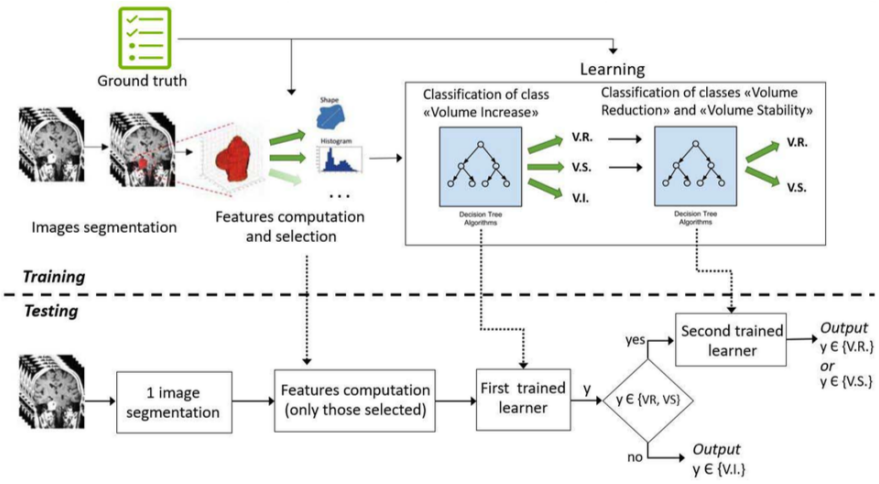


Fig. B.4 Schematic representation of the proposed machine learning approach.

Rosa Sicilia

(3) Reference Sicilia, R., Cordelli, E., Merone, M., Luperto, E., Papalia, R., Iannello, G., and Soda, P. (2019, June). Early radiomic experiences in classifying prostate cancer aggressiveness using 3D local binary patterns. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 355-360). IEEE.

Abstract Prostate cancer is the most common form of cancer in Western countries and there is the need to develop clinical decision support systems able to support physicians in the diagnosis of clinical relevant prostate cancer and avoid useless invasive prostate biopsies. In this respect, this paper introduces a radiomic approach that classifies the prostate cancer aggressiveness by combining Three Orthogonal Planes-Local Binary Pattern (TOP-LBP) with other texture measures. Furthermore, to combat the skewed nature of class priors, our proposal employs a data augmentation technique (Figure B.5). The results achieved on 99 samples are up-and-coming, they favorably compare against conventional PI- RADS-based approach, and they show also the benefit given by the introduction of TOP-LBP in the radiomic signature, reaching an accuracy of 0.82.

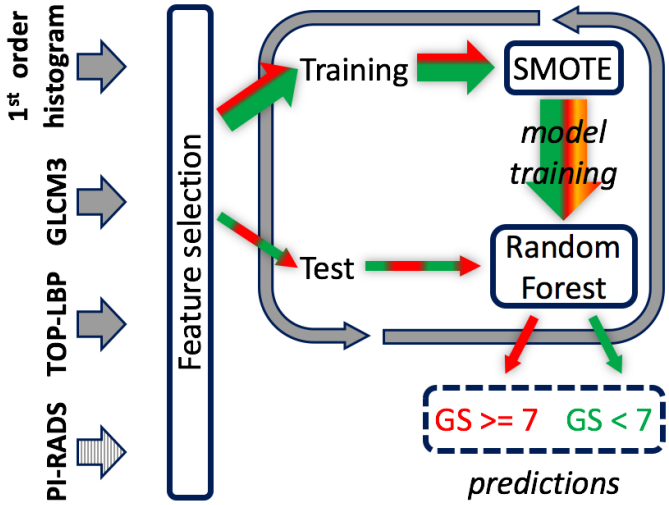


Fig. B.5 Schematic representation of the proposed approach.

Rosa Sicilia