Gaussian Processes for Rumour Stance Classification in Social Media

MICHAL LUKASIK and KALINA BONTCHEVA, University of Sheffield, United Kingdom
TREVOR COHN, University of Melbourne, Australia
ARKAITZ ZUBIAGA, University of Warwick, United Kingdom
MARIA LIAKATA and ROB PROCTER, University of Warwick and Alan Turing Institute, United Kingdom

Social media tend to be rife with rumours while new reports are released piecemeal during breaking news. Interestingly, one can mine multiple reactions expressed by social media users in those situations, exploring their stance towards rumours, ultimately enabling the flagging of highly disputed rumours as being potentially false. In this work, we set out to develop an automated, supervised classifier that uses multi-task learning to classify the stance expressed in each individual tweet in a conversation around a rumour as either supporting, denying or questioning the rumour. Using a Gaussian Process classifier, and exploring its effectiveness on two datasets with very different characteristics and varying distributions of stances, we show that our approach consistently outperforms competitive baseline classifiers. Our classifier is especially effective in estimating the distribution of different types of stance associated with a given rumour, which we set forth as a desired characteristic for a rumour-tracking system that will show both ordinary users of Twitter and professional news practitioners how others orient to the disputed veracity of a rumour, with the final aim of establishing its actual truth value.

CCS Concepts: • **Information systems** → **Social networks**; **Clustering and classification**; *Data stream mining*;

Additional Key Words and Phrases: Social media, rumours, stance classification, veracity classification, breaking news, machine learning

ACM Reference format:
Michal Lukasik, Kalina Bontcheva, Trevor Cohn, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2019. Gaussian Processes for Rumour Stance Classification in Social Media. *ACM Trans. Inf. Syst.* 37, 2, Article 20 (February 2019), 24 pages.
https://doi.org/10.1145/3295823

Michal Lukasik is now at Google.

This work is partially supported by the European Union under grant agreement No. 611233 PHEME and No. 654024 (SoBigData). This research utilised Queen Mary's MidPlus computational facilities, supported by QMUL Research-IT and funded by EPSRC grant EP/K000128/1.

Authors' addresses: M. Lukasik and K. Bontcheva, University of Sheffield, Sheffield, United Kingdom; emails: lukasikmic@gmail.com, k.bontcheva@sheffield.ac.uk; T. Cohn, University of Melbourne, Melbourne, Australia; email: t.cohn@unimelb.edu.au; A. Zubiaga, M. Liakata, and R. Procter, University of Warwick, Coventry, United Kingdom; emails: {a.zubiaga, m.liakata, rob.procter}@warwick.ac.uk.

1046-8188/2019/02-ART20
https://doi.org/10.1145/3295823

1 INTRODUCTION

There is an increasing need to interpret and act upon rumours spreading quickly through social media during breaking news, where new reports are released piecemeal and often have an unverified status at the time of posting. Previous research has posited the damage that the diffusion of false rumours can cause in society, and that corrections issued by news organisations or state agencies such as the police may not necessarily achieve the desired effect sufficiently quickly [17, 34]. Being able to determine the accuracy of reports is therefore crucial in these scenarios. However, the veracity of rumours in circulation is usually hard to establish [2], since as many views and testimonies as possible need to be assembled and examined to reach a final judgement. Examples of rumours that were later disproven, after being widely circulated, include a 2010 earthquake in Chile, where rumours of a volcano eruption and a tsunami warning in Valparaiso spawned on Twitter [26]. Another example is the England riots in 2011 (see Table 1), where false rumours claimed that rioters were going to attack the Birmingham Children's Hospital and that animals had escaped from the London Zoo [35].

Previous work by ourselves and others has argued that looking at how users in social media orient to rumours is a crucial first step towards making an informed judgement on the veracity of a rumourous report [26, 41, 42, 48]. For example, in the case of the riots in England in August 2011, Procter et al. manually analysed the stance expressed by users on social media towards rumours [35]. Using a coding frame devised for the purpose, each tweet discussing a rumour was manually categorised as supporting, denying or questioning it. It is obvious that manual methods have their disadvantages in that they do not scale well; the ability to perform stance categorisation of tweets in an automated way would be of great use in tracking rumours and subsequently determining their likely veracity, for instance flagging those that are largely denied or questioned as being more likely to be false.

Determining the stance of social media posts automatically has been attracting increasing interest in the scientific community in recent years, as this is a useful first step towards more in-depth rumour analysis. For example, previous work has used stance classification systems to determine the veracity of rumours [19]. In this case, a stance classifier described earlier in Reference [36] was used, with the addition of new rule-based features; however, the evaluation was performed on the veracity classification task, and did not assess the performance of the stance classifier. In other research, the detection of questioning tweets, which would be one of the outcomes of a stance classification system, has been used to classify an event as a rumour or a non-rumour [45]. However, their approach for detecting questioning tweets relies on manually defined regular expressions and thus can hardly be generalised to new events [47]. Instead, we argue that the automated approach to rumour stance classification that we advocate can help to achieve this.

Work on automatic rumour stance classification, however, is still in its infancy, with some approaches assuming an unrealistic evaluation scenario, conducting cross-validation rather than splitting the tweets so that a classifier is tested strictly on future tweets (e.g., Reference [36]). Our work advances the state-of-the-art in tweet-level stance classification through multi-task learning and Gaussian Processes. This article substantially extends our earlier short paper [22], first by using a second dataset, which enables us to test the generalisability of our results. Second, a comparison against additional baseline classifiers and recent state-of-the-art approaches has been added to the experimental section. Finally, we carried out a more thorough analysis of the results, now including per-class performance scores as well as an entropy-based analysis of features, which furthers our understanding of rumour stance classification.

Table 1. Tweets Pertaining to a Rumour about Hospital Being Attacked during 2011 England Riots

Text	Position
Birmingham Children's hospital has been attacked. F***ing morons. #UKRiots	support
Girlfriend has just called her ward in Birmingham Children's Hospital & there's no sign of any trouble #Birminghamriots	deny
Birmingham children's hospital guarded by police? Really? Who would target a childrens hospital #disgusting #Birminghamriots	question

In comparison to the current state-of-the-art, our approach is novel in several crucial aspects:

(1) We perform stance classification on *unseen rumours*, given a training set of already annotated rumours on different topics and from different time periods. In addition, we run experiments with a small initial number of tweets from the target rumour being available for the classifier during training and evaluating it on the future tweets.
(2) *Generalisability to new datasets* is a core aspect of our methodology, which is built on the premise that patterns of stance should exhibit similar characteristics across different rumours.
(3) *Application of Gaussian Processes with multi-task learning kernels*, which are state-of-the-art in many NLP tasks [4, 6, 16, 24, 33], however, have not been applied to rumour stance classification before. We demonstrate how this model achieves superior results to frequentist baselines and how the multi-task learning kernels help achieve the aforementioned generalisability across multiple rumours.

Based on the assumption of a common underlying linguistic signal in rumours on different topics, we build a transfer learning system based on Gaussian Processes, which can classify stance in tweets discussing newly emerging rumours. The article reports results on two different rumour datasets and explores two different experimental settings—without any training data and with very limited training data. We refer to these as follows:

- *Leave One Out (LOO)*: all tweets pertaining to a target rumour are only used for testing, i.e., method performance on a completely unseen rumour is reported;
- *Leave Part Out (LPO)*: the first few tweets of a target rumour (as annotated by journalists) and added to the training set of the Gaussian Process classifier, together with tweets pertaining to older rumours. The rest of the tweets on the target rumour are used for evaluation.

Both experimental settings are transfer learning scenarios, where reference training rumours are available for training. They differ in that the LOO setting poses a problem where no annotation for the test rumour is available, whereas in LPO, a few initial tweets are additionally used for training, thus making for a potentially easier problem, while admittedly requiring an additional (small) annotation effort.

Our results demonstrate that a Gaussian Process-based transfer learning approach leads to significantly improved performance over the Gaussian Process-based single task learning, and competitive results compared to the state-of-the-art methods and competitive baselines, as demonstrated on two very different datasets. The classifier relying on Gaussian Processes performs particularly well over the rest of the baseline classifiers in the Leave Part Out setting, proving that it does particularly well in determining the distribution of supporting, denying and questioning tweets associated with a rumour. Estimating the distribution of stances is the key aspect for which our classifier performs especially well compared to the baseline classifiers.

2 RELATED WORK

This section provides a more in-depth motivation of the rumour stance detection task and an overview of the state-of-the-art methods and their limitations. First, however, let us start by introducing the formal definition of a rumour.

2.1 Rumour Definition

There have been multiple attempts at defining rumours in the literature. Most of them are complementary to one another, with slight variations depending on context. The core concept on which most researchers agree matches the definition that major dictionaries provide, such as the Oxford English Dictionary[1] defining a rumour as *"a currently circulating story or report of uncertain or doubtful truth."* For instance, the authors of Reference [8] defined rumours as "unverified and instrumentally relevant information statements in circulation."

Researchers have long looked at the properties of rumours to understand their diffusion patterns and to distinguish them from other kinds of information that people habitually share [9]. Allport and Postman [2] claimed that rumours spread due to two factors: people want to find meaning in events and, when faced with ambiguity, people resort to finding meaning through telling stories. The latter also explains why rumours tend to change in time by becoming shorter, sharper and more coherent. This is the case, it is argued, because in this way rumours explain things more clearly. However, Rosnow [38] claimed that there are four important factors for rumour transmission: rumours must be outcome-relevant to the listener; must increase personal anxiety; be somewhat credible; and be uncertain. Furthermore, Shibutani [39] defined rumours to be *"a recurrent form of communication through which men [sic] caught together in an ambiguous situation attempt to construct a meaningful interpretation of it by pooling their intellectual resources. It might be regarded as a form of collective problem-solving."*

In contrast with these three theories, Guerin and Miyazaki [12] state that a rumour is a form of relationship-enhancing talk. Building on their previous work, they argue that kinds of talk serve the purpose of forming and maintaining social relationships. Rumours, they say, can be explained by such means.

In our work, we adhere to the widely accepted fact that rumours are unverified pieces of information. More specifically, following Zubiaga et al. [48], we define a rumour in the context of breaking news, as a *"circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety so as to motivate finding out the actual truth."*

2.2 Descriptive Analysis of Rumours in Social Media

One particularly influential piece of work in the field of rumour analysis in social media is that by Mendoza et al. [26]. By manually analysing the data from the earthquake in Chile in 2010, the authors selected seven confirmed truths and seven false rumours, each consisting of close to 1,000 tweets or more. The veracity value of the selected stories was corroborated by using reliable sources. Each tweet from each of the news items was manually classified into one of the following classes: affirmation, denial, questioning, unknown or unrelated. In this way, each tweet was classified according to the orientation its author displayed towards the topic it was about. The study showed that a much higher percentage of tweets about false rumours are shown to deny the respective rumours (approximately 50%). This is in contrast to rumours later proven to be true, where only 0.3% of tweets were denials. Based on this, authors claimed that rumours can be detected using aggregate analysis of the stance expressed in tweets.

Recent research published in a special issue on rumours and social media [31] also shows the increasing interest of the scientific community in the topic. Webb et al. [43] proposed an agenda

[1] http://www.oxforddictionaries.com/definition/english/rumour.

for research that establishes an interdisciplinary methodology to explore in full the propagation and regulation of unverified content in social media. Middleton and Krivcovs [27] described an approach for geoparsing social media posts in real-time, which can be of help to determine the veracity of rumours by tracking down the poster's location. The contribution of Hamdi et al. [13] to rumour resolution is to build an automated system that rates the level of trust of users in social media, hence enabling the elimination of users with low reputation. Complementary to these approaches, our objective is to determine the stance of tweets towards a rumour, which can then be aggregated to establish an overall veracity score for the rumour.

Another study that shows insightful conclusions with respect to stance towards rumours is that by Procter et al. [35]. The authors conducted an analysis of a large dataset of tweets related to the riots in the UK, which took place in August 2011. The dataset collected in the riots study is one of the two used in our experiments, and we describe it in more detail in Section 3.4. After grouping the tweets into topics, where each topic represents a rumour, they were manually categorised into different classes, namely (1) media reports, which are tweets sent by mainstream media accounts or journalists connected to media; (2) pictures, being tweets with a link to images; (3) rumours, being tweets claiming or counter claiming something without giving any source; and (4) reactions, consisting of responses of users to the riots phenomenon or specific event related to the riots. Besides categorisation of tweets by type, Procter et al. also manually categorised the accounts posting tweets into different types, such as mainstream media, only on-line media, activists, celebrities, bots, among others. What is interesting for the purposes of our work is that the authors observed the following four-step pattern recurrently occurring across the collected rumours: (1) a rumour is initiated by someone claiming it may be true, (2) a rumour spreads together with its reformulations, (3) counter claims appear, and (4) a consensus emerges about the credibility of the rumour.

This leads the authors to the conclusion that the process of "inter-subjective sense making" by Twitter users plays a key role in exposing false rumours. This finding, together with subsequent work by Tolmie et al. into the conversational characteristics of microblogging [41] has motivated our research into automating stance classification as a methodology for accelerating this process.

2.3 Rumour Stance Classification

Qazvinian et al. [36] conducted early work on rumour stance classification. They introduced a system that analyses a set of tweets associated with a given topic predefined by the user. Their system would then classify each of the tweets as supporting, denying, or questioning a tweet. We have adopted this scheme in terms of the different types of stance in the work we report here. However, their work ended up merging denying and questioning tweets for each rumour into a single class, converting it into a two-way classification problem of supporting vs denying-or-questioning. Instead, we keep those classes separate and, following Procter et al. [34], we conduct a three-way classification [49].

Another important aspect that differentiates Qazvinian et al.'s work from ours is that they looked at support and denial on longstanding rumours, such as the fact that many people conjecture whether or not Barack Obama is a Muslim. By contrast, we look at rumours that emerge in the context of fast-paced, breaking news situations, where new information is released piecemeal, often with statements that employ hedging words such as "reportedly" or "according to sources" to make it clear that the information is not fully verified at the time of posting. This is a very different scenario from that in Qazvinian et al.'s work, as the emergence of rumourous reports can lead to sudden changes in vocabulary, leading to situations that might not have been observed in the training data.

Another aspect that we deal with differently in our work, aiming to make it more realistically applicable to a real-world scenario, is that we apply the method to each rumour separately.

Ultimately, our goal is to classify new, emerging rumours, which can differ from what the classifier has observed in the training set. Previous work ignored this separation of rumours, by pooling together tweets from all the rumours in their collections, both in training and test data. By contrast, we consider the rumour stance classification problem as a form of transfer learning and seek to classify unseen rumours by training the classifier from previously labelled rumours. We argue that this makes a more realistic classification scenario towards implementing a real-world rumour-tracking system.

Following a short interlude, there has been a burst of renewed interest in this task since 2015. For example, Liu et al. [19] introduced rule-based methods for stance classification, which were shown to outperform the approach by Qazvinian et al. [36]. Similarly, Zhao et al. [45] use regular expressions instead of an automated method for identifying enquiring tweets that question the veracity of a rumour; a rumour stance classification may help enhance their approach, which did not seek to identify other kinds of responses. Hamidian and Diab [14] used Tweet Latent Vectors to assess the ability of performing two-way classification of the stance of tweets as either supporting or denying a rumour. They investigated the extent to which a model trained on historical tweets could be used for classifying new tweets on the same rumour. This, however, limits the method's applicability to long-running rumours only.

The work closest to ours in terms of aims is that of Zeng et al. [44], who explored the use of three different classifiers for automated rumour stance classification on unseen rumours. In their study, classifiers were set up on a two-way classification problem dealing with tweets that support or deny rumours. In the present work, we extend this research by performing three-way classification that also deals with tweets that question the rumours. Moreover, we adopt the three classifiers used in their work, namely Random Forest, Naive Bayes, and Logistic Regression, as baselines in our work.

Zhao et al. [45] focused on the related task of rumour detection by first detecting enquiring tweets, i.e., tweets that challenge the accuracy of a post. However, the authors did not perform stance classification as such but, instead, manually defined regular expressions to look for enquiring tweets. The broader rumour stance classification task, in an automated way that can generalise to new rumours, was not explored by these authors. In contrast, we target the problem of stance classification within rumours, which allows us to make an assumption of a common underlying characteristic of data, supported by a hypothesis that rumours exhibit similar characteristics [35]. Thus, gathering data coming from rumours only makes for a reasonable approach.

Stance classification is a problem that occurs also in non-rumour applications, with examples of political leaning classification [46] and debate stance classification [40] for detection of agreement and disagreement. Rumour stance classification is different from these applications in that it is most beneficial to conduct it early during the rumour spread, so that appropriate officials may react quickly. This motivates our settings where little or no annotation per rumour is available. Other characteristic of rumours that may not be present in other applications is sparsity of available data and a multitudeness of rumours around major events as shown in the case of our datasets. Moreover, similar patterns are exhibited by different rumours [35], which supports the hypothesis of a common underlying signal.

3 PROBLEM DEFINITION: TWEET LEVEL RUMOUR STANCE CLASSIFICATION

3.1 Definition of the Task

Individual tweets may discuss the same rumour in different ways, where each tweet author expresses their own stance towards the rumour. Within this scenario, we define the tweet level rumour stance classification task as that in which a classifier has to determine the stance of each

tweet towards the rumour. More specifically, given the tweet t_i as input, the classifier has to determine which one of the set $Y = \{supporting, denying, questioning\}$ applies to the tweet, $y(t_i) \in Y$. Further, we define the task as a supervised classification problem, where the classifier is trained from a labelled set of tweets and is applied to tweets on a new, unseen set of rumours.

3.2 Definition of the Stances

Below, we define the three different stances that a tweet can take with respect to a rumour.

Supporting. A supporting tweet unambiguously expresses or suggests a belief that a rumour is true. It can provide a support for a rumour either by linking a supposedly factual content (a picture, a link to a story), by providing a description of how the tweet author or their acquaintances witnessed the rumour or by explaining why the story seems credible. Support can also be expressed simply by the author expressing the feelings that the rumour triggered.

Denying. A denying tweet unambiguously expresses a disbelief in a rumour. It can undermine its credibility, explain why its author thinks it is not credible. A tweet can provide any kind of evidence of a similar nature to that listed in the supporting case, e.g., links to websites debunking a rumour, witnessing stories that undermine a rumour.

Questioning. A questioning tweet may support or challenge the veracity of a rumour, but because it does so in an ambiguous manner, for example, by asking for more information, it does not belong in either of the previous categories. Questioning tweets can often be replies to supporting or denying tweets.

3.3 Problem Formulation

Let $R = \{R_1, \ldots, R_{|R|}\}$ be a set of rumours, each of which consists of posts (tweets) discussing it, $\forall_{m=1,\ldots,|R|}\, R_m = \{\mathbf{p}_m^1, \ldots, \mathbf{p}_m^{|R_m|}\}$. $P = \cup_{m=1,\ldots,|R|} R_m$ is the complete set of tweets from all rumours. Each tweet is classified as supporting, denying or questioning with respect to its rumour: $\forall_{\mathbf{p}\in P}\, y(\mathbf{p}) \in Y = \{\text{supporting, denying, questioning}\}$.

Previous work evaluated the rumour stance classification task using cross-validation. In this approach, the set of all tweets P is randomly split among K folds (Reference [36] used $K = 5$), and iteratively each fold is used as a test set, and the remaining $K - 1$ folds serve as a training set. In Figure 1(a), we show an illustration of one fold in this setting, with question marks denoting tweets from the test set and other symbols denoting labels from the training set. Notice how training tweets occur after the test tweets within the same rumour, a scenario that does not occur in real-world settings where journalists are interested in obtaining stances expressed in the most recent tweets. Ultimately, the separation of rumours and time dependencies were ignored in evaluation of previous work. Here, we deal with the task differently, arguing that the evaluation from previous work does not correspond to a real-world scenario. In applications one should be able to classify new, emerging rumours, which can differ from what the classifier has observed in the training set.

We formulate the problem in settings that better reflect the real-world scenario. First, we consider the LOO setting, in which for each rumour $R_m \in R$ we construct the test set equal to R_m and the training set equal to $P \setminus R_m$. This is the most challenging scenario, where the test set contains an entirely unseen rumour. We depict it in Figure 1(b). Thus, we apply the method to each rumour separately. Ultimately, we consider the rumour stance classification problem as a form of transfer learning and seek to classify unseen rumours by training the classifier on previously annotated rumours. We argue that this makes for a more realistic classification scenario towards implementing a real-world rumour-tracking system.

The second setting is LPO. Here, a number of initial tweets from the target rumour R_m is added to the training set $\{\mathbf{p}_m^1, \ldots, \mathbf{p}_m^k\}$, as depicted in Figure 1(c). This scenario becomes applicable typically

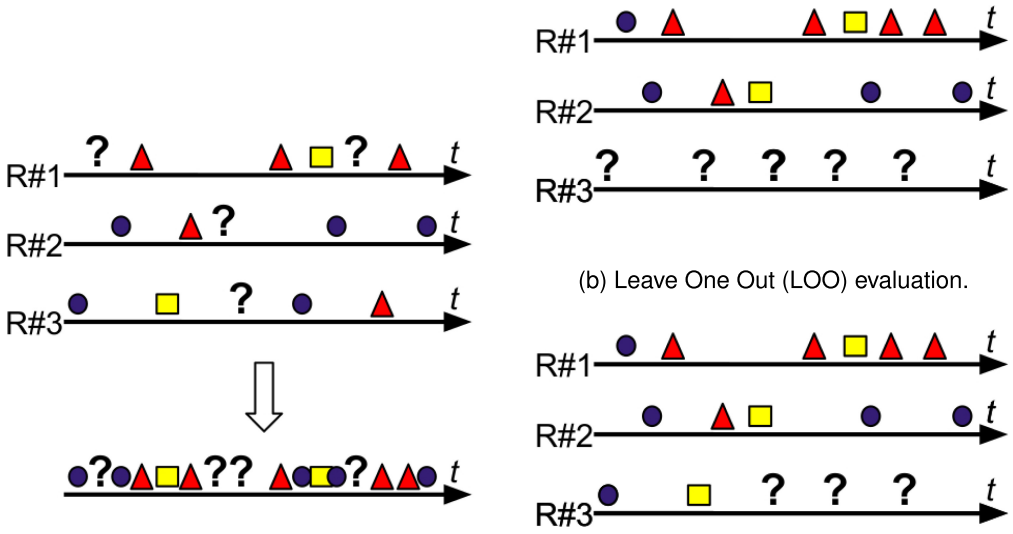


Fig. 1. Illustration of different evaluation techniques for rumour stance classification. Different symbols correspond to tweets from one of the rumours that occurred at a specific point of time. Question marks denote the tweets that need to be classified in the test phase, other symbols denote observed classes of tweets in the training set (blue circles denote supporting tweets, yellow squares denote questioning tweets, red triangles denote rejecting tweets). The evaluation from previous work ignores rumour identities and time dependencies between tweets (depicted in the left subfigure), conflating all rumours into one (shown at the bottom line). In our approach, we focus on predicting labels for tweets from a left-out rumour strictly into the future (depicted in the right subfigures).

soon after a rumour breaks out and journalists have started monitoring and analysing the related tweet stream. In our experiments, we consider $k \in \{10, 20, 30, 40, 50\}$.

Notice that in these settings future tweets can still be present in the training set as long as they come from reference (non-test) rumours, and as such are not strictly realistic. The riot events we consider are short-lived, with rumours of short lifespans. This results in rumours overlapping in time, and so keeping only non-overlapping past rumours would result in very little reference data being kept for training. Therefore, here we keep reference rumours regardless of when they occurred.

The tweet-level stance classification problem here assumes that tweets from the training set are already categorized into what rumours they discuss. This information can be acquired either via manual annotation as part of initial analysis by journalists, as is the case with our dataset, or automatically, e.g., using pattern-based rumour detection [45]. Our method is then used to classify the stance expressed in each new tweet from the test set.

3.4 Datasets

We evaluate our work on two different datasets, which we describe below. We use two recent datasets from previous work for our study, both of which adapt to our needs. We do not use the dataset by Reference [36] given that it uses a different annotation scheme limited to two categories of stances. The reason we do not combine them is that they have very different characteristics, and

Table 2. Counts of Tweets with Supporting, Denying, or Questioning Labels in Each Rumour Collection from the England Riots Dataset

Rumour	Supporting	Denying	Questioning
Army bank	62	42	73
Children's hospital	796	487	132
London Eye	177	295	160
McDonald's	177	0	13
Miss Selfridge's	3150	0	7
Police beat girl	783	4	95
London zoo	616	129	99
Total	5761	957	579

in this way our approach enables us to assess the ability of our classifier to deal with these different characteristics.

3.4.1 England Riots Dataset. The first dataset consists of several rumours circulating on Twitter during the England riots in 2011 (see Table 2). The dataset was collected by tracking a long set of keywords associated with the event. The dataset was analysed and annotated manually as supporting, questioning, or denying a rumour, by a team of social scientists studying the role of social media during the riots [35].

As can be seen from the dataset overview in Table 2, different rumours exhibit varying proportions of supporting, denying and questioning tweets, which was also observed in other studies of rumours [26, 36]. These variations in the number of instances for each class across rumours poses a challenge for properly modelling a rumour stance classifier. The classifier needs to be able to deal with a test set where the distribution of classes can be very different to that observed in the training set.

Thus, we perform sevenfold cross-validation in the experiments, each fold having six rumours in the training set, and the remaining rumour in the test set. The seven rumours are as follows [35]:

- Rioters had attacked London Zoo and released the animals.
- Rioters were gathering to attack Birmingham's Children's Hospital.
- Rioters had set the London Eye on fire.
- Police had beaten a sixteen year old girl.
- The Army was being mobilised in London to deal with the rioters.
- Rioters had broken into a McDonalds and set about cooking their own food.
- A store belonging to the Miss Selfridge retail group had been set on fire in Manchester.

3.4.2 PHEME Dataset. In addition, we use another rumour dataset associated with five different events, which was collected as part of the PHEME FP7 research project and described in detail in References [48, 50]. Note that the authors released datasets for nine events, but here we remove non-English datasets, as well as small English datasets each of which includes only 1 rumour, as opposed to the 40+ rumours in each of the datasets that we are using. We summarise the details of the five events we use from this dataset in Table 3.

In contrast to the England riots dataset, the PHEME datasets were collected by tracking conversations initiated by rumourous tweets. This was done in two steps. First, we collected tweets that contained a set of keywords associated with a story unfolding in the news. We will be referring

Table 3. Counts of Tweets with Supporting, Denying, or Questioning Labels in Each Event Collection on the PHEME Dataset

Dataset	Rumours	Supporting	Denying	Questioning
Ottawa shooting	58	161	76	63
Ferguson riots	46	192	82	94
Charlie Hebdo	74	235	56	51
Germanwings crash	68	67	12	28
Sydney siege	71	222	89	99
Total	287	877	315	335

to the latter as an event. Next, we sampled the most retweeted tweets, on the basis that rumours, by definition, should be "a circulating story that produces sufficient skepticism or anxiety." This allows us to filter potentially rumourous tweets and collect the "conversational threads" [41] initiated by them. The threads were tracked by collecting replies to tweets and, therefore, unlike the England riots, by definition this dataset also comprises replying tweets. This is an important characteristic of the dataset, as one would expect that replies are generally shorter and potentially less descriptive than the source tweets that initiated the conversation. We take this difference into consideration when performing the analysis of our results.

This dataset includes tweets associated with the following five events:

- **Ferguson unrest:** Citizens of Ferguson (USA) protested after the fatal shooting of an 18-year-old African American, Michael Brown, by a white police officer on August 9, 2014.
- **Ottawa shooting:** Shootings occurred on Ottawa's Parliament Hill in Canada, resulting in the death of a Canadian soldier on October 22, 2014.
- **Sydney siege:** A gunman held as hostages ten customers and eight employees of a Lindt chocolate café located at Martin Place in Sydney, Australia, on December 15, 2014.
- **Charlie Hebdo shooting:** Two brothers forced their way into the offices of the French satirical weekly newspaper Charlie Hebdo in Paris, killing 11 people and wounding 11 more, on January 7, 2015.
- **Germanwings plane crash:** A passenger plane from Barcelona to Düsseldorf crashed in the French Alps on March 24, 2015, killing all passengers and crew on board. The plane was ultimately found to have been deliberately crashed by the co-pilot of the plane.

In this case, we perform fivefold cross-validation, having four events in the training set and the remaining event in the test set for each fold.

4 EXPERIMENT SETTINGS

This section details the features and evaluation measures used in our experiments on tweet level rumour stance classification.

4.1 Classifiers

We begin by describing the classifiers we use for our experimentation, including Gaussian Processes, as well as a set of competitive baseline classifiers that we use for comparison.

4.1.1 Gaussian Processes for Classification. Gaussian Processes are a Bayesian non-parametric machine learning framework that has been shown to work well for a range of NLP problems, often beating other state-of-the-art methods [4, 6, 16, 21, 24, 33].

A Gaussian Process defines a prior over functions, which combined with the likelihood of data points gives rise to a posterior over functions explaining the data. The key concept is a kernel function, which specifies how outputs correlate as a function of the input. Thus, from a practitioner's point of view, a key step is to choose an appropriate kernel function capturing the similarities between inputs to optimise the way the problem is modelled.

Gaussian Processes exhibit many useful properties that make them appealing for rumour stance classification. First, this probabilistic kernelised framework avoids the need for expensive cross-validation for hyperparameter selection.[2] Instead, the marginal likelihood of the data can be used for hyperparameter selection. Moreover, it provides information about the uncertainty of the classifications, which allows a user to decide whether a prediction is reliable. Even though we do not use this information in this work, we demonstrate that GPs work better than the baselines, and in applications the uncertainty information could be used as an additional source of information.

The central concept of Gaussian Process Classification (GPC; Reference [37]) is a latent function f over inputs $\mathbf{x}$: $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where m is the mean function, assumed to be 0, and k is the kernel function, specifying the degree to which the outputs covary as a function of the inputs. We use a linear kernel, $k(\mathbf{x}, \mathbf{x}') = \sigma^2 \mathbf{x}^\top \mathbf{x}'$. The latent function is then mapped by the probit function $\Phi(f)$ into the range $[0, 1]$, such that the resulting value can be interpreted as $p(y = 1|\mathbf{x})$.

The GPC posterior is calculated as

$$p(f^*|X, \mathbf{y}, \mathbf{x}_*) = \int p(f^*|X, \mathbf{x}_*, \mathbf{f}) \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y}|X)} d\mathbf{f},$$

where $p(\mathbf{y}|\mathbf{f}) = \prod_{j=1}^{n} \Phi(f_j)^{y_j}(1 - \Phi(f_j))^{1-y_j}$ is the Bernoulli likelihood of class y. After calculating the above posterior from the training data, this is used in prediction, i.e.,

$$p(y_* = 1|X, \mathbf{y}, \mathbf{x}_*) = \int \Phi(f_*)\, p(f_*|X, \mathbf{y}, \mathbf{x}_*)\, df_*.$$

The above integrals are intractable and approximation techniques are required to solve them. There exist various methods to deal with calculating the posterior; here we use Expectation Propagation (EP; Reference [29]). In EP, the posterior is approximated by a fully factorised distribution, where each component is assumed to be an unnormalised Gaussian.

To conduct multi-class classification, we perform a one-vs.-all classification for each label and then assign the one with the highest likelihood amongst the three (supporting, denying, questioning). We choose this method due to interpretability of results, similar to recent work on occupational class classification [33].

Transfer Learning. In the LPO setting initial labelled tweets from the target rumour are observed as well, as opposed to the LOO setting. In the case of LPO, we propose to weigh the importance of tweets from the reference rumours based on how similar their characteristics are to the tweets from the target rumour available for training. To handle this with GPC, we use a multiple output model based on the Intrinsic Coregionalisation Model (ICM; Reference [3]). This model has already been applied successfully to NLP regression problems [4] and it can also be applied to classification ones. ICM parametrises the kernel by a matrix that represents the extent of covariance between pairs of tasks. The complete kernel takes form of

$$k((\mathbf{x}, d), (\mathbf{x}', d')) = k_{data}(\mathbf{x}, \mathbf{x}')B_{d,d'},$$

where B is a square coregionalisation matrix, d and d' denote the tasks of the two inputs and k_{data} is a kernel for comparing inputs $\mathbf{x}$ and $\mathbf{x}'$ (here, linear). Thus, the similarity function between the

[2]There exist frequentist kernel methods, such as SVMs, which additionally require extensive heldout parameter tuning.

two tweets is a product of inter-rumour similarity ($B_{d,d'}$) and a tweet similarity independent from the rumour identities (k_{data}). This allows to conduct transfer learning by weighting the importance of annotated tweets from the reference training rumours based on how similar the characteristics of the reference rumours are to that of a test rumour. We parametrise the coregionalisation matrix $B = \boldsymbol{\kappa} I + \boldsymbol{v}\boldsymbol{v}^T$, where $\boldsymbol{v}$ specifies the correlation between tasks and the vector κ controls the extent of task independence. Note that in case of LOO setting this model does not provide useful information, since no target rumour data is available to estimate similarity with respect to other rumours.

Hyperparameter Selection. We tune hyperparameters $\mathbf{v}$, $\boldsymbol{\kappa}$ and σ^2 by maximising evidence of the model $p(\mathbf{y}|X)$, thus having no need for a validation set.

Methods. We consider GPs in three settings, varying in what data the model is trained on and what kernel it uses:

GP Only Target considers only target rumour data for training, and thus only uses the single task learning kernel. Notice that this model setting can not be considered in the LOO problem setting, as it would not have access to any training data.

GP considers both the target rumour data as well as the reference rumours data (i.e., other than the target rumour), however, only uses the single task learning kernel.

GP-ICM considers both the target rumour data as well as the reference rumours data (i.e., other than the target rumour), and employs the ICM kernel for learning the similarities between different rumours.

4.1.2 Baselines. To assess and compare the efficiency of Gaussian Processes for rumour stance classification, we also experimented with baseline classifiers:

Majority vote classifier based on the training label distribution.

Logistic Regression (MaxEnt) was the first method employed for rumour stance classification [36]. We use ℓ_1 regularisation with the cost coefficient selected from the list: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100]. The cost coefficient was found using grid search employing threefold cross-validation over the training set, where two folds were used for training and one for evaluation of the proposal coefficient.

Support Vector Machines (SVM) with the cost coefficient selected via nested cross-validation from the list of values: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100]. The cost coefficient was found using grid search employing threefold cross-validation over the training set.

Random Forests (RF) which [44] found to be the best approach in their experiments with rumour stance classification. The authors report the value of only one hyperparameter value, namely they set the number of trees to 30, although they do not state whether it is chosen via hyperparameter optimization. A Random Forests classifier is controlled by a number of hyperparameters, which we select via grid search over the cross product between the considered hyperparameter values (employing threefold cross-validation over the training set). The hyperparameters that we consider are as follows: the splitting criterion measuring the quality of a split (optimized for from the list [Gini impurity, entropy]), the number of trees (optimized for from the list [10, 50, 100, 150, 200]), the minimum number of samples in a node to perform a split (optimized for from the list [2, 5, 10]). We use bootstrap samples when choosing data for each tree.

We use Scikit-learn implementations of the baseline classifiers [32].

4.2 Features

We conducted a series of preprocessing steps to address data sparsity. All words were converted to lowercase, stopwords have been removed,[3] all emoticons were replaced by words,[4] and stemming was performed. In addition, multiple occurrences of a character were replaced with a double occurrence [1], to correct for misspellings and lengthening, e.g., *looool*. All punctuation was also removed, except for ., *!* and *?*, which we hypothesise to be important for expressing emotion. Lastly, usernames were removed as they tend to be rumour specific, i.e., very few users comment on more than one rumour.

After preprocessing the text data, we use either the resulting bag of words (BOW) feature representation and replace all words with their Brown cluster ids (Brown). Brown clustering is a hard hierarchical clustering method [18]. It clusters words based on maximising the probability of the words under the bigram language model, where words are generated based on their clusters. In previous work it has been shown that Brown clusters yield better performance than directly using the BOW features [22].

In our experiments, the clusters used were obtained using 1,000 clusters acquired from a large scale Twitter corpus [30], from which we can learn Brown clusters aimed at representing a generalisable Twitter vocabulary. Retweets are removed from the training set to prevent bias [20]. More details on the Brown clusters that we used as well as the words that are part of each cluster are available online.[5]

During the experimentation process, we also tested additional features, including the use of the bag of words instead of the Brown clusters, as well as using word embeddings obtained from the training sets [28]. However, the results turned out to be substantially poorer than those we obtained with the Brown clusters. We conjecture that this was due to the little data available to train the word embeddings; further exploring use of word embeddings trained from larger training datasets is left for future work.

4.3 Evaluation Measures

Accuracy is often deemed a suitable evaluation measure to assess the performance of a classifier on a multi-class classification task. However, the classes are clearly imbalanced in our case, with varying tendencies towards one of the classes in each of the rumours. We argue that in these scenarios the sole evaluation based on accuracy is insufficient, and further measurement is needed to account for category imbalance. This is especially necessary in our case, as a classifier that always predicts the majority class in an imbalanced dataset will achieve high accuracy, even if the classifier is useless in practice. To tackle this, we use both micro-averaged and macro-averaged F1 scores. Note that the micro-averaged F1 score is equivalent to the well-known accuracy measure, while the macro-averaged F1 score complements it by measuring performance assigning the same weight to each category.

Both of the measures rely on precision (Equation (1)) and recall (Equation (2)) to compute the final F1 score,

$$\text{Precision}_k = \frac{tp_k}{tp_k + fp_k}, \tag{1}$$

$$\text{Recall}_k = \frac{tp_k}{tp_k + fn_k}, \tag{2}$$

[3]We removed stopwords using the English list from Python's NLTK package.
[4]We used the dictionary from: http://bit.ly/1rX1Hdk and extended it with: :o, : |, =/, :s, :S, :p.
[5]http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html.

where tp_k (true positives) refer to the number of instances correctly classified in class k, fp_k (false positives) is the number of instances incorrectly classified in class k, and fn_k (false negatives) is the number of instances that actually belong to class k but were not classified as such.

The above equations can be used to compute precision and recall for a specific class. Precision and recall for all the classes in a problem with c classes are computed differently if they are microaveraged (see Equations (3) and (4)) or macroaveraged (see Equations (5) and (6)),

$$\text{Precision}_{\text{micro}} = \frac{\sum_{k=1}^{c} tp_k}{\sum_{k=1}^{c} tp_k + \sum_{k=1}^{c} fp_k}, \tag{3}$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{k=1}^{c} tp_k}{\sum_{k=1}^{c} tp_k + \sum_{k=1}^{c} fn_k}, \tag{4}$$

$$\text{Precision}_{\text{macro}} = \frac{\sum_{k=1}^{c} \text{Precision}_k}{c}, \tag{5}$$

$$\text{Recall}_{\text{macro}} = \frac{\sum_{k=1}^{c} \text{Recall}_k}{c}. \tag{6}$$

After computing microaveraged and macroveraged precision and recall, the final F1 score is computed in the same way, i.e., calculating the harmonic mean of the precision and recall in question (see Equation (7)),

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{7}$$

After computing the F1 score for each fold, we compute the micro-averaged score across folds.

5 RESULTS

In this section, we report results of experiments on rumour stance classification. We analyze the performance of different methods and draw conclusions about the task. We report Micro-F1 and Macro-F1 metrics that provide insights about two aspects of model performance: how well it classifies tweets overall (i.e., minimizing the absolute number of errors) and how well it balances the errors for different stances.

Tables 4, 5, and 6 report the combined results in both the LOO and LPO settings. Consecutive columns correspond to an increasing number of tweets from the target rumour available during training (column 0 corresponds to the LOO setting, and other columns correspond to the LPO setting). In Table 4, we show results for the GP-based models using different text representations. Notice that in the case of both England riots and PHEME datasets, Brown clusters make for a more robust text representation. Brown clusters always yield better results on the PHEME dataset according to both Micro-F1 and Macro-F1 scores. Moreover, on the England riots dataset, Brown clusters always lead to a better Macro-F1 score, and to competitive Micro-F1 scores. Thus, in the following analysis we report baselines using the more promising text representation employing Brown clusters. We discuss the relative performance of different GP settings in more detail in the following sections.

5.1 Experiments on the England Riots Dataset

In Table 5, we report micro-averaged and macro-averaged F1 scores of methods' performance on the England riots dataset as the number of tweets from the target rumour used for training increases (this information is graphically illustrated in Figure 2). Notice how performance of *GP Only Target* is significantly lower than that of *GP* and *GP-ICM*, showing the importance of using additional data from reference rumours. We can notice that the performance of most of the

Table 4. Micro-F1 and Macro-F1 Scores for GP-based Methods under Different Settings Using Different Word Representation Methods (Brown Clusters and BOW; Denoted by Rows) and Different Proportions of the Initial Tweets Annotated from the Target Rumour/Event on the England Riots and the PHEME Datasets (Denoted by Columns)

		0	10	20	30	40	50
Macro-F1	GP Only Target Brown	N/A	0.346	0.366	0.366	0.382	0.416
	GP Only Target BOW	N/A	0.314	0.346	0.379	0.388	0.402
	GP Brown	**0.489**	0.571	0.620	0.614	0.615	0.617
	GP BOW	0.452	0.572	0.603	0.593	0.588	0.616
	GP-ICM Brown	0.436	**0.634**	**0.708**	**0.646**	**0.657**	**0.635**
	GP-ICM BOW	0.394	0.510	0.585	0.544	0.574	0.562
Micro-F1	GP Only Target Brown	N/A	0.787	0.722	0.733	0.735	0.769
	GP Only Target BOW	N/A	0.781	0.710	0.760	0.751	0.775
	GP Brown	0.614	0.737	0.765	0.761	0.762	0.763
	GP BOW	**0.640**	0.817	0.825	0.818	0.811	**0.833**
	GP-ICM Brown	0.540	0.812	**0.855**	**0.829**	**0.833**	0.828
	GP-ICM BOW	0.476	**0.821**	0.806	0.809	0.811	0.822

(a) England riots

		0	10	20	30	40	50
Macro-F1	GP Only Target Brown	N/A	0.434	0.489	0.494	0.514	0.515
	GP Only Target BOW	N/A	0.356	0.382	0.399	0.415	0.439
	GP Brown	0.548	**0.555**	0.569	0.566	0.567	0.575
	GP BOW	0.465	0.472	0.475	0.477	0.471	0.481
	GP-ICM Brown	**0.557**	**0.555**	**0.592**	**0.575**	**0.594**	**0.598**
	GP-ICM BOW	0.453	0.465	0.455	0.439	0.466	0.471
Micro-F1	GP Only Target Brown	N/A	0.546	0.577	0.612	0.606	0.613
	GP Only Target BOW	N/A	0.591	0.548	0.554	0.546	0.558
	GP Brown	0.631	**0.636**	0.644	0.644	0.645	0.650
	GP BOW	0.551	0.569	0.572	0.575	0.572	0.579
	GP-ICM Brown	**0.655**	0.635	**0.652**	**0.655**	**0.668**	**0.675**
	GP-ICM BOW	0.561	0.579	0.587	0.578	0.580	0.577

(b) PHEME

methods improves as the proportion of annotated training examples from the target rumour increases. This phenomenon is especially noticeable for the GP-ICM method. Notice that when no annotation from the target rumour is used, its performance is poor in terms of micro-averaged F1 score. However, it is able to make very effective use of the annotation. Its performance keeps improving as the number of training instances approaches 50 and overtakes the baselines after 20 annotated examples. This shows GP-ICM is able to make use of the labelled instances from the target rumour, which the baselines struggle with. Note that 50 tweets represent, on average, less than 7% of the whole rumour, with the rest of the rumour unobserved during training. Moreover, notice how regardless of the number of labelled instances, GP-ICM yields good results in terms of macro-averaged F1 score. This shows that GP-ICM balances between the errors made for each stance better than other models. Last, we notice that SVM achieves competitive results that are above the rest of the baselines, outperforming GP on both metrics. Notice that only GP-ICM and

Table 5. Micro-F1 and Macro-F1 Scores for Different Methods on the England Riots Dataset

		0	10	20	30	40	50
Macro-F1	Majority	0.294	0.294	0.294	0.294	0.294	0.294
	GP Only Target	N/A	0.346	0.366	0.366	0.382	0.416
	GP	0.489	0.571	0.620	0.614	0.615	0.617
	GP-ICM	0.436	**0.634**	**0.708**	**0.646**	**0.657**	**0.635**
	MaxEnt	0.491	0.529	0.569	0.611	0.575	0.577
	SVM	**0.535**	0.614	0.632	0.626	0.629	0.629
	RF	0.491	0.514	0.522	0.531	0.510	0.526
Micro-F1	Majority	**0.788**	0.788	0.788	0.788	0.788	0.788
	GP Only Target	N/A	0.787	0.722	0.733	0.735	0.769
	GP	0.614	0.737	0.765	0.761	0.762	0.763
	GP-ICM	0.540	**0.812**	**0.855**	**0.829**	**0.833**	**0.828**
	MaxEnt	0.633	0.658	0.720	0.774	0.759	0.761
	SVM	0.701	0.794	0.808	0.805	0.806	0.808
	RF	0.757	0.771	0.775	0.790	0.775	0.784

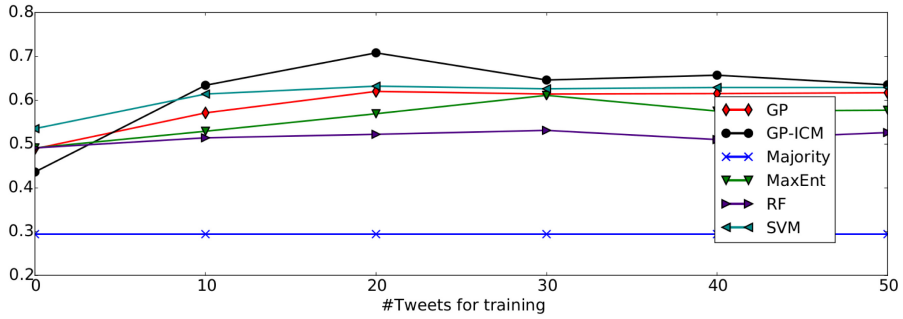
Table 6. Micro-F1 and Macro-F1 Scores for Different Methods on the PHEME Dataset

		0	10	20	30	40	50
Macro-F1	Majority	0.240	0.240	0.240	0.240	0.240	0.240
	GP Only Target	N/A	0.434	0.489	0.494	0.514	0.515
	GP	0.548	0.555	0.569	0.566	0.567	0.575
	GP-ICM	0.557	0.555	0.592	0.575	0.594	0.598
	MaxEnt	0.544	0.544	0.551	0.549	0.555	0.559
	SVM	0.590	0.590	0.589	0.594	0.591	0.591
	RF	**0.593**	**0.599**	**0.606**	**0.611**	**0.604**	**0.609**
Micro-F1	Majority	0.561	0.561	0.561	0.561	0.561	0.561
	GP Only Target	N/A	0.546	0.577	0.612	0.606	0.613
	GP	0.631	0.636	0.644	0.644	0.645	0.650
	GP-ICM	0.655	0.635	0.652	0.655	0.668	0.675
	MaxEnt	0.649	0.648	0.653	0.653	0.652	0.655
	SVM	0.677	0.678	0.678	0.681	0.680	0.680
	RF	**0.692**	**0.694**	**0.698**	**0.702**	**0.696**	**0.702**

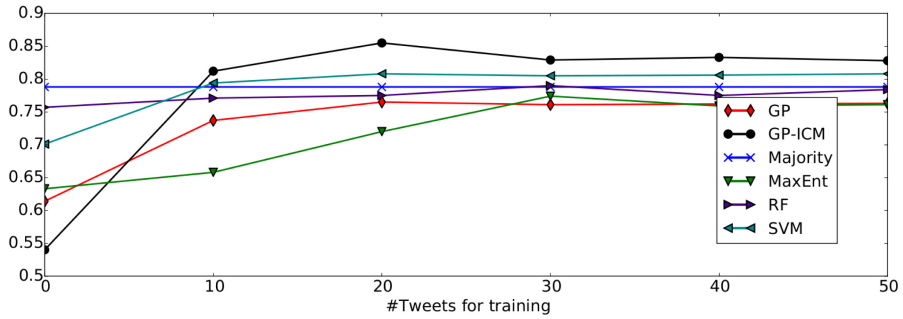
SVM are able to consistently beat the Majority classifier once 20 tweets from the training rumour are available for training.

5.2 Experiments on the PHEME Dataset

In Table 6, we report micro-averaged and macro-averaged F1 scores of methods' performance on the PHEME dataset as the number of tweets from the target rumour used for training increases (this information is graphically illustrated in Figure 3). Notice the results are overall lower than in the case of the England Riots dataset. The reason for this is twofold. First, in the PHEME dataset we deal with a more challenging setting, where at each fold of the evaluation we leave out an event



(a) Macro-F1



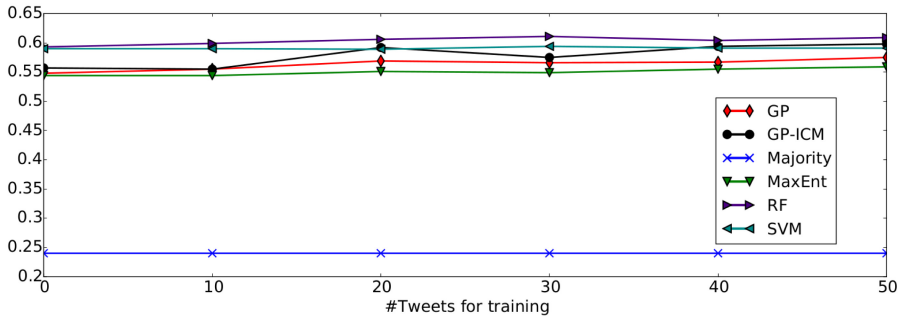
(b) Micro-F1

Fig. 2. Macro-F1 and Micro-F1 scores for different methods over the number of tweets from the target rumour used for training on the England riots dataset. The test set is fixed to all but the first 50 tweets of the target rumour, making the results comparable across the varying training size.

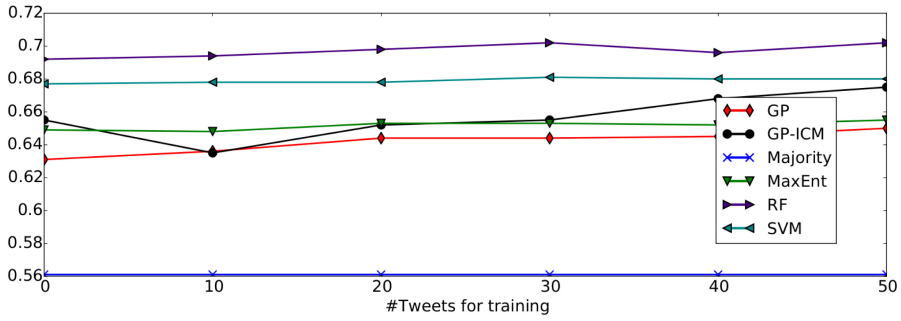
out (where an event is composed of rumours), and train on other events. Instead, in the England riots case we were leaving one rumour out within the same event. Secondly, the PHEME dataset is largely composed of tweets that are replying to others [48], which makes them shorter. Moreover, rumours from the PHEME dataset are much shorter than rumours from the England riots dataset, and hence more challenging to get meaningful features from. Despite these difficulties, we are interested in exploring if similar trends hold across classifiers.

One difference from the England Riots results is that, in this case, the classifiers are not benefiting as much from incorporating increasing number of annotated tweets from the target rumour. This is likely due to the heterogeneity of the events from the PHEME dataset. Namely, within each event there are multiple rumours, and as the number of initial tweets is annotated, we are gaining insight into a diverse set of rumours as they start to unfold. All of these rumours are different, and they are not necessarily covering all rumours from the target event. By contrast, each left out set of tweets in the England Riots pertains to a single rumour, and so annotating its initial tweets is useful, giving insights into characteristics about that rumour.

We observe that Random Forests turn out to be the best approach according to both metrics. Interestingly, the second best method is different depending on which metric we consider. According to Macro-F1, GP-ICM and SVM are both competitive, with GP-ICM being slightly better with larger supervision. However, under Micro-F1, SVM is clearly the second best approach. Similarly



(a) Macro-F1



(b) Micro-F1

Fig. 3. Macro-F1 and Micro-F1 scores for different methods over the number of tweets from the target rumour used for training on the England riots and the PHEME datasets. The test set is fixed to all but the first 50 tweets of the target rumour, making the results comparable across the varying training size.

to the England Riots results, the performance of *GP Only Target* is significantly lower than that of *GP* and *GP-ICM*, showing that reference rumour annotations is crucial for the GPs to achieve competitive results (we omit *GP Only Target* from the graphs for better visualization of differences between the best performing models). Moreover, *GP-ICM* outperforms *GP*, which shows that the multi-task learning kernel brings improvements within the same model.

5.3 Analysis of the Best Performing Methods

Next, we analyze the results of the best-performing classifiers (GP-ICM, SVM, and RF) by looking at the per-class performance. Tables 7 and 8 report per-class F1 scores for the three best performing classifiers for the England riots dataset and the PHEME dataset in LPO settings where 20 tweets from a target rumour are available during training. The table also reports statistics on the misclassifications that the approaches made (the cross-stance classifications are also depicted graphically for the case of England riots in Figure 4).

Notice that in the case of the England riots, GP-ICM is a clear winner. It is the only approach that manages to retrieve more than 50% of denies, which is one of the two under-represented stances (denies are around 6 times less frequent than supports, whereas questions are around 10 times less frequent than supports). Interestingly, the questioning stance is easier to correctly classify than the denying stance across the methods, even though it is even less frequent.

Table 7. Per-class Precision, Recall and F1 Scores for the Best-performing Classifiers on the England Riots Dataset

Class	Classifier	Performance			Deviations		
		P	R	F1	S	D	Q
Supporting (S)	GP-ICM	**0.893**	**0.935**	**0.914**	0.935	0.008	0.057
	RF	0.833	0.903	0.867	0.903	0.014	0.082
	SVM	0.873	0.896	0.884	0.896	0.019	0.086
Denying (D)	GP-ICM	**0.882**	**0.535**	**0.666**	0.452	0.535	0.013
	RF	0.584	0.162	0.254	0.823	0.162	0.015
	SVM	0.742	0.423	0.539	0.563	0.423	0.014
Questioning (Q)	GP-ICM	**0.496**	**0.602**	**0.544**	0.352	0.045	0.602
	RF	0.380	0.540	0.446	0.403	0.057	0.540
	SVM	0.394	0.593	0.473	0.339	0.068	0.593

Table 8. Per-class Precision, Recall, and F1 Scores for the Best-performing Classifiers on the PHEME Dataset

Class	Classifier	Performance			Deviations		
		P	R	F1	S	D	Q
Supporting (S)	GP-ICM	**0.748**	0.767	0.757	0.767	0.133	0.101
	RF	0.714	**0.899**	**0.796**	0.899	0.053	0.047
	SVM	0.706	0.865	0.777	0.865	0.071	0.064
Denying (D)	GP-ICM	0.442	**0.385**	**0.412**	0.428	0.385	0.187
	RF	**0.570**	0.277	0.373	0.594	0.277	0.129
	SVM	0.511	0.259	0.344	0.604	0.259	0.137
Questioning (Q)	GP-ICM	0.588	**0.625**	0.606	0.233	0.141	0.625
	RF	**0.708**	0.601	**0.650**	0.329	0.071	0.601
	SVM	0.676	0.618	0.646	0.318	0.064	0.618

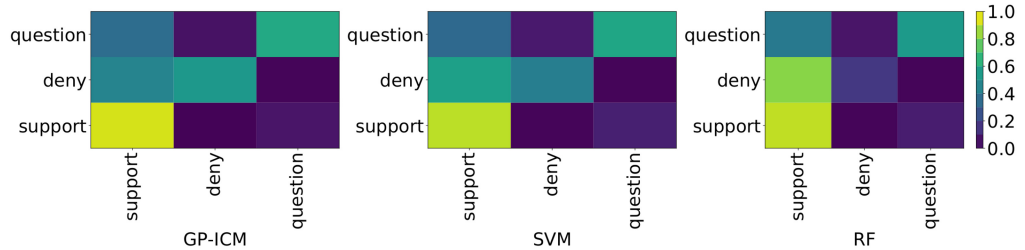


Fig. 4. Cross-classification rates for competitive methods on the England riots dataset. A cell i, j denotes what percentage of times the ground truth stance i is being classified as stance j. The statistics are also reported in Table 7.

In the case of the PHEME rumours, GP-ICM is again the best classifier in terms of retrieving the denies. This is an interesting property, as denying is the most challenging stance (challenging in the sense of yielding the worst results across all methods). However, for both supporting and questioning stances RF and SVM make better predictions.

Table 9. Most Representative Keywords for the Three Categories in the England Riots Dataset

Supporting	Denying	Questioning
rt	like	release
control	eye	bank
apparently	claim	army
actual	yeah	last
escape	n	hope
go	though	aint
tiger	untrue	londoney
hear	think	incident
seen	story	prayforlondon
see	would	tbh

The problem of misclassifying denies is due to the datasets' imbalance, which is a common problem in the rumour stance distribution, as previous studies have shown that users rarely deny or question rumours but instead largely support rumours regardless of whether they are true or false [48].

We ran Wilcoxon signed rank test for evaluating statistical significance of differences between GP-ICM and GP performance. To evaluate significance of improvements of GP-ICM over GP, for each dataset we apply the Wilcoxon test to all the tweets across rumours, where the values in the two sequences denote the correctness of the predicted label for a tweet. We found that GP-ICM is better than GP on the England riots dataset for all cases where any target rumour supervision is available (i.e., for the supervision from the target rumour equal to 10, 20, 30, 40, 50) with $p < 0.05$. On the PHEME dataset, we did not find statistically significant difference between GP-ICM and GP.

5.4 Feature Analysis

To conclude the analysis, and to better understand the features that best characterise the three types of stances, we perform a feature analysis for the two datasets under study. For this analysis, we make use of the Kullback-Leibler divergence [15] to compute the keywords with highest likelihood for each stance using an entropy-based metric. In Tables 9 and 10, we list the top 10 keywords for each of the three labels for the England riots and PHEME datasets.

We observe some interesting patterns that characterise and are indicative of the labels. This is the case for the *Denying* category, with negation words such as *no* or *nothing*, negative conjunctions like *though*, or direct references to fakeness with words like *untrue*. The *Supporting* category, on the other hand, has a number of keywords from people tweeting claiming to have witnessed the event in some way, such as *hear*, *seen* or *see*, keywords pointing to potential information sources such as *read* or *press*, or specific attributions to people directly involved in the event, such as *killer* or *found*. Finally, the *Questioning* category is more diverse for not having a clear supporting or denying stance, but still shows that people tweeting compose questions with keywords such as *anyone* or *anything*, show confusion (*confused*) or desperation (*please* or *smh*).

5.5 Discussion

We experimented with two rumour datasets, and adapted the introduced evaluation schemes differently to each of them. In the case of the first dataset, we were making predictions for single held out rumours from the England riots event from 2011, thus dealing with the setting where all

Table 10. Most Representative Keywords for the Three Categories in the PHEME Dataset

Supporting	Denying	Questioning
action	help	make
around	would	got
even	three	confused
prime	no	anything
vigilant	run	hear
killer	know	anyone
found	time	xu
indicate	gun	sveb
read	nothing	please
press	signal	smh

the rumours are revolving around the same background event. In the case of the second dataset, we were making predictions for each of the five different events, having access to the remaining four, making for a significantly more challenging setup. In these different settings, we made various observations regarding the relative performance of different approaches. We observed that while a GP trained only on the target data is not achieving competitive results, the GP using reference examples (the scenario of all other baselines) achieves better performance, and its multi-task learning variant GP-ICM leads to additional improvements leading to outperforming the baselines in the case of the England riots dataset. Moreover, in the case of both datasets, GP-ICM manages to perform relatively well in classifying the denying stance, which turns out to be the most challenging. Another appealing aspect of GP-ICM in the England riots dataset is that it performs well despite having very few annotations from the target rumour, making better use of such training data than the baselines. However, we notice that when annotation comes from external events (which is the case for the PHEME dataset experiments), Random Forests and SVMs are competitive with the GP-ICM approach. This poses a question of whether multi-task learning variants of Random Forests or SVMs [10] could bring further improvements, as we found the multi-task learning Gaussian process model (GP-ICM) to consistently outperform the single-task learning GP across all settings.

6 CONCLUSIONS

Social media is becoming an increasingly important tool for maintaining social resilience: Individuals use it to express opinions and follow events as they unfold; news media organisations use it as a source to inform their coverage of these events; and government agencies, such as the emergency services, use it to gather intelligence to help in decision-making and in advising the public about how they should respond [34]. While previous research has suggested that mechanisms for exposing false rumours are implicit in the ways in which people use social media [35], it is nevertheless critically important to explore if there are ways in which computational tools can help to accelerate these mechanisms so that misinformation and disinformation can be targeted more rapidly and the benefits of social media to society maintained [7].

As a first step towards achieving this aim, this article has investigated the problem of classifying the different types of stance expressed by individuals in tweets about rumours. First, we considered a setting where no training data from the target rumours is available (LOO). Without access to annotated examples of the target rumour the learning problem becomes very difficult. We showed that in the supervised domain adaptation setting (LPO), even annotating a small number of tweets

helps to achieve better results. Moreover, we demonstrated the benefits of a multi-task learning approach, as well as that Brown cluster features are more useful for the task than simple bag of words.

Findings from previous work, such as Castillo et al. [5] and Procter et al. [35], suggested that the aggregate stance of individual users tends to correlate with actual rumour veracity. Hence, the next step in our own work will be to make use of the classifier for the stance expressed in the reactions of individual Twitter users to determine the actual veracity of the rumour in question. Another interesting direction for future work is the addition of non-textual features to the classifier. For example, rumour diffusion patterns [23] have been applied to rumour stance classification in the follow-up work [25].

ACKNOWLEDGMENTS

The work was implemented using the GPy toolkit [11].

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media (LSM'11)*. Association for Computational Linguistics, 30–38.

[2] G. W. Allport and L. Postman. 1947. The psychology of rumor. *J. Clin. Psychol.* (1947). https://psycnet.apa.org/record/1948-00288-000.

[3] Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. 2012. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.* 4, 3 (2012), 195–266.

[4] Daniel Beck, Trevor Cohn, and Lucia Specia. 2014. Joint emotion analysis via multi-task Gaussian processes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1798–1803.

[5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Res.* 23, 5 (2013), 560–588.

[6] Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*. 32–42.

[7] Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, Sara-Jayne Terp, Geraldine Wong, Christian Burger, Arkaitz Zubiaga, Rob Procter, and Maria Liakata. 2015. PHEME: Computing veracity the fourth challenge of big social data. In *European Semantic Web Conference ESWC*. 25–29.

[8] Nicholas DiFonzo and Prashant Bordia. 2007. Rumor, gossip and urban legends. *Diogenes* 54, 1 (2007), 19–35.

[9] Pamela Donovan. 2007. How idle is idle talk? One hundred years of rumor research. *Diogenes* 54, 1 (2007), 59–82.

[10] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. 109–117.

[11] The GPy authors. 2015. GPy: A Gaussian process framework in Python. Retrieved from http://github.com/SheffieldML/GPy.

[12] Bernard Guerin and Yoshihiko Miyazaki. 2006. Analyzing rumors, gossip, and urban legends through their conversational properties. *Psychol. Rec.* 56, 1, Article 2 (2006).

[13] Sana Hamdi, Alda Lopes Gancarski, Amel Bouzeghoub, and Sadok Ben Yahia. 2016. TISoN: Trust inference in trust-oriented social networks. *ACM Trans. Inf. Syst.* 34, 3 (2016), 17.

[14] Sardar Hamidian and Mona T. Diab. 2016. Rumor identification and belief investigation on Twitter. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*. 3–8.

[15] Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *Ann. Math. Stat.* 22, 1 (1951), 79–86.

[16] Vasileios Lampos, Nikolaos Aletras, Daniel Preotiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*. 405–413.

[17] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction continued influence and successful debiasing. *Psychol. Sci. Publ. Interest* 13, 3 (2012), 106–131.

[18] Percy Liang. 2005. *Semi-Supervised Learning for Natural Language*. Ph.D. Dissertation. Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology.

[19] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on Twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM'15)*. ACM, New York, NY, 1867–1870. DOI: https://doi.org/10.1145/2806416.2806651

[20] Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (26-31) *(LREC'14)*. 462–468.

[21] Michal Lukasik and Trevor Cohn. 2016. Convolution kernels for discriminative learning from streaming text. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*.

[22] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Classifying tweet level judgements of rumours in social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. 2590–2595.

[23] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Point process modelling of rumour dynamics in social media. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL'15)*. 518–523.

[24] Michal Lukasik, P. K. Srijith, Trevor Cohn, and Kalina Bontcheva. 2015. Modeling tweet arrival times using log-Gaussian Cox processes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. 250–255.

[25] Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: An application to rumour stance classification in Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*. 393–398.

[26] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we RT? In *Proceedings of the 1st Workshop on Social Media Analytics (SOMA'10)*. 71–79.

[27] Stuart E. Middleton and Vadims Krivcovs. 2016. Geoparsing and geosemantics for social media: Spatio-temporal grounding of content propagating rumours to support trust and veracity analysis during breaking news. *ACM Trans. Inf. Syst.* 34, 3 (2016), 1–27.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[29] Thomas Minka and John Lafferty. 2002. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI'02)*. 352–359.

[30] Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'13)*. 380–390.

[31] Symeon Papadopoulos, Kalina Bontcheva, Eva Jaho, Mihai Lupu, and Carlos Castillo. 2016. Overview of the special issue on trust and veracity of information in social media. *ACM Trans. Inf. Syst.* 34, 3 (2016), 14.

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.

[33] Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL'15)*. 1754–1764. http://aclweb.org/anthology/P/P15/P15-1169.pdf.

[34] Rob Procter, Jeremy Crump, Susanne Karstedt, Alex Voss, and Marta Cantijoch. 2013. Reading the riots: What were the police doing on Twitter? *Pol. Soc.* 23, 4 (2013), 413–436.

[35] Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on Twitter: Methodological innovation for the analysis of big data. *Int. J. Soc. Res. Methodol.* 16, 3 (2013), 197–214.

[36] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. 1589–1599.

[37] Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[38] Ralph L. Rosnow. 1991. The psychology of rumor. *Am. Psychol.* 46, 5 (1991), 484–496.

[39] Tamotsu Shibutani. 1969. Improvised news: A sociological study of rumor. *Soc. Res.* 36, 1 (1969), 169–171.

[40] Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*.

[41] Peter Tolmie, Rob Procter, Mark Rouncefield, Maria Liakata, and Arkaitz Zubiaga. 2015. Microblog analysis as a programme of work. *arXiv preprint arXiv:1511.03193* (2015).

[42] Peter Tolmie, Rob Procter, Mark Rouncefield, Maria Liakata, Arkaitz Zubiaga, and Dave Randall. 2017. Supporting the use of user generated content in journalistic practice. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

[43] Helena Webb, Pete Burnap, Rob Procter, Omer Rana, Bernd Carsten Stahl, Matthew Williams, William Housley, Adam Edwards, and Marina Jirotka. 2016. Digital wildfires: Propagation, verification, regulation, and responsible innovation. *ACM Trans. Inf. Syst.* 34, 3 (2016), 15.

[44] Li Zeng, Kate Starbird, and Emma S. Spiro. 2016. # unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*.

[45] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Early detection of rumors in social media from enquiry posts. In *Proceedings of the International World Wide Web Conference Committee (IW3C2)*.

[46] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. 2011. Classifying the political leaning of news articles and users from user votes. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'11)*. 417–424.

[47] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363* (2016).

[48] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE* 11, 3 (03 2016), 1–29. DOI: https://doi.org/10.1371/journal.pone.0150989

[49] Arkaitz Zubiaga, Peter Tolmie, Maria Liakata, and Rob Procter. 2014. D2.1 development of an annotation scheme for social media rumours. *PHEME Deliverable* (2014). http://www.pheme.eu/wp-content/uploads/2016/02/PHEME-D2.1-Development_of_an_annotation_scheme.pdf.

[50] Arkaitz Zubiaga, Peter Tolmie, Maria Liakata, and Rob Procter. 2015. *D2.4 Qualitative Analysis of Rumours, Sources, and Diffusers Across Media and Languages*. Technical Report. University of Warwick.

Received August 2016; revised October 2018; accepted November 2018