Fine-grained Topic Detection and Tracking on Twitter

Nicholas Mamo[a], Joel Azzopardi and Colin Layfield[b]
Faculty of ICT, University of Malta, Malta

Keywords: Twitter, Topic Detection and Tracking, Information Retrieval.

Abstract: With its large volume of data and free access to information, Twitter revolutionised Topic Detection and Tracking (TDT). Thanks to Twitter, TDT could build timelines of real-world events in real-time. However, over the years TDT struggled to adapt to Twitter's noise. While TDT's solutions stifled noise, they also kept the area from building granular timelines of events, and today, TDT still relies on large datasets from popular events. In this paper, we detail Event TimeLine Detection (ELD) as a solution: a real-time system that combines TDT's two broad approaches, document-pivot and feature-pivot methods. In ELD, an on-line document-pivot technique clusters a stream of tweets, and a novel feature-pivot algorithm filters clusters and identifies topical keywords. This mixture allows ELD to overcome the technical limitations of traditional TDT algorithms to build fine-grained timelines of both popular and unpopular events. Nevertheless, our results emphasize the importance of robust topic tracking and the ability to filter subjective content.

1 INTRODUCTION

In 2006, Twitter's launch transformed real-world events into social experiences, but it also transformed the roles of its users. Twitter gave regular individuals the role of amateur reporters, and the platform itself became an informal newswire. Twitter's users generate hundreds of thousands of tweets, even during relatively unpopular events (Meladianos et al., 2015). With so much information, Topic Detection and Tracking (TDT) went from reproducing what news outlets had reported to discovering the news for itself. Modern techniques have been applied to identify breaking news (Cataldi et al., 2013), but they can also go further and build timelines of specific events (Saeed et al., 2019b).

The TDT community explored these tasks, but while Twitter's user base grew over time, TDT's performance did not improve. Like older research (Zhao et al., 2011), modern works can recognize changes in tweeting behaviour as important developments, like the spike in tweets after a goal in a football match, shown in Figure 1. However, machines scarcely do better. Even today, non-key topics, like yellow cards in football matches, continue to elude algorithms (Meladianos et al., 2015; Gillani et al., 2017).

There are several obstacles in TDT's way. Some of the blame lies in Twitter's informal and brief tweets, but TDT algorithms are also at fault. In this paper, we describe Event TimeLine Detection (ELD), outlined previously (Mamo et al., 2019), as a way of easing these challenges. ELD, a TDT approach based on FIRE: Finding Important News REports (FIRE) (Mamo and Azzopardi, 2017), combines document-pivot and feature-pivot techniques to build more detailed timelines, understand topics better and suppress noise. We make the following contributions:

- We previously used FIRE to detect breaking news. In this paper, we experiment with the combination of document-pivot and feature-pivot methods in specified, or planned, events.
- FIRE batches tweets before processing them, delaying reporting. ELD replaces tweet batching with a snapshot procedure to operate in real-time.
- ELD includes a novel feature-pivot algorithm. Our technique uses a new burst metric to give more context to topics by extracting the keywords that describe them.

ELD's contributions show how TDT can build better timelines in spite of Twitter's difficulties. The rest of this paper is structured as follows. We discuss TDT's difficulties on Twitter in Section 2, and we describe ELD's design in Section 3. In Section 4, we present an analysis on six football matches, and show how ELD builds more granular timelines than tradi-

[a] https://orcid.org/0000-0001-5452-5548
[b] https://orcid.org/0000-0002-1868-4258

Mamo, N., Azzopardi, J. and Layfield, C.
Fine-grained Topic Detection and Tracking on Twitter.
DOI: 10.5220/0010639600003064
In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021) - Volume 1: KDIR*, pages 79-86
ISBN: 978-989-758-533-3; ISSN: 2184-3228

tional methods. We conclude in Section 5.

2 RELATED WORK

In TDT research, there is a time before Twitter and a time after Twitter. Before Twitter, TDT would wait for news outlets to publish articles and then group them into events, tracking them until they expired (Farzindar and Khreich, 2015). After Twitter launched in 2006, its API gave TDT access to a large volume of real-time data and the possibility to chase breaking news (Saeed et al., 2019a).

Although TDT has improved since 2006, it failed to capitalize on Twitter's data. In football matches, for example, algorithms identify key topics, like goals, but miss non-key topics, like yellow cards (Meladianos et al., 2015; Gillani et al., 2017). Non-key topics generally evoke less passion than key topics (Meladianos et al., 2015), but the lack of enthusiasm does not explain why so many algorithms miss non-key topics even from enormous datasets. We identify three reasons to explain TDT's challenges.

First, TDT algorithms themselves are limited. Traditionally, TDT was a clustering, or document-pivot, problem. TDT harnessed years of research on clustering, but it also inherited clustering's challenges, namely fragmentation (Aiello et al., 2013). Later, feature-pivot techniques sought to solve some of document-pivot models' problems. Usually, feature-pivot techniques look for bursts in the number of published documents (Zhao et al., 2011), or more intense use of some related keywords (Cataldi et al., 2013). The intuition behind burst is that when something momentous happens, like a goal in a football match, discussion changes (Meladianos et al., 2015).

However, burst-based methods contributed new challenges. Bursty keywords are less expressive than documents in a cluster, which communicate the subject clearly (Aiello et al., 2013). Many algorithms cluster terms to form a topic (Cataldi et al., 2013), but a group of terms can still be misleading (Aiello et al., 2013). In addition, methods based on burst tend to let key topics dwarf non-key topics occurring close to each other (Saeed et al., 2019b), such as a goal followed by a substitution.

Second, Twitter is a difficult medium. Twitter's tweets, no longer than 280 characters, complicate any Information Retrieval (IR) task, including TDT. The short length limits the amount of information in tweets (Unankard et al., 2015). Brevity, in particular, is detrimental to document-pivot methods, which rely on comparisons among tweets for clustering. Brevity weakens even established methods, like the Term Frequency-Inverse Document Freqeuncy (TF-IDF) scheme. Since words rarely repeat in the same tweet, TF-IDF's Inverse Document Frequency (IDF) component becomes an inverse term frequency. Consequently, some argue that TF-IDF is not designed for short content (Unankard et al., 2015).

Moreover, the volume and velocity at which tweets are published restrict the amount of processing that algorithms can perform. For example, many document-pivot techniques avoid complex and time-consuming clustering approaches in favour of on-line algorithms (Mamo and Azzopardi, 2017).

Third, Twitter users and journalists write differently. Twitter users tweet informally and with little regard to orthography, unlike reporters. And while an article becomes newsworthy as soon as the media publishes it, tweets are an outlet for users to share opinions, react emotionally, and narrate their everyday life (Farzindar and Khreich, 2015). Identifying and removing noise is a major challenge in TDT, especially in events like football matches, which are conducive to passionate outbursts. Most techniques have a low tolerance for noise. For example, it is common to remove all retweets because they introduce bias (Saeed et al., 2019b). Although such excessive filtering may seem sensible, retweets are an integral part of Twitter's conversations.

Excessive filtering stifles TDT's potential. Generally, document-pivot techniques accept the largest clusters, or clusters whose size exceeds a threshold, as topics (Ifrim et al., 2014). Nevertheless, the more aggressive the filtering, the fewer tweets remain, and consequently fewer clusters reach the threshold. TDT's caution on Twitter is justifiable, but in popular events, aggressive filtering inhibits algorithms from detecting both key and non-key topics. In unpopular events, the number of tweets is too small to filter aggressively. As a result, almost all research focuses on popular events (Saeed et al., 2019a).

In this paper, we build on FIRE (Mamo and Azzopardi, 2017) to overcome some of Twitter's challenges. Contrary to most TDT approaches, FIRE combines document-pivot and feature-pivot techniques (Mamo and Azzopardi, 2017). FIRE bins tweets into five-minute, non-overlapping time windows, and clusters tweets to identify potential topics. The algorithm considers clusters with as few as three tweets, but it filters them using a neural network and, more importantly, a feature-pivot approach. FIRE uses Cataldi et al. (2013)'s algorithm to compare how words in the cluster had been used in the past and how they are being used presently. The algorithm only accepts clusters if their central words had not been popular recently. In other words, FIRE's feature-pivot

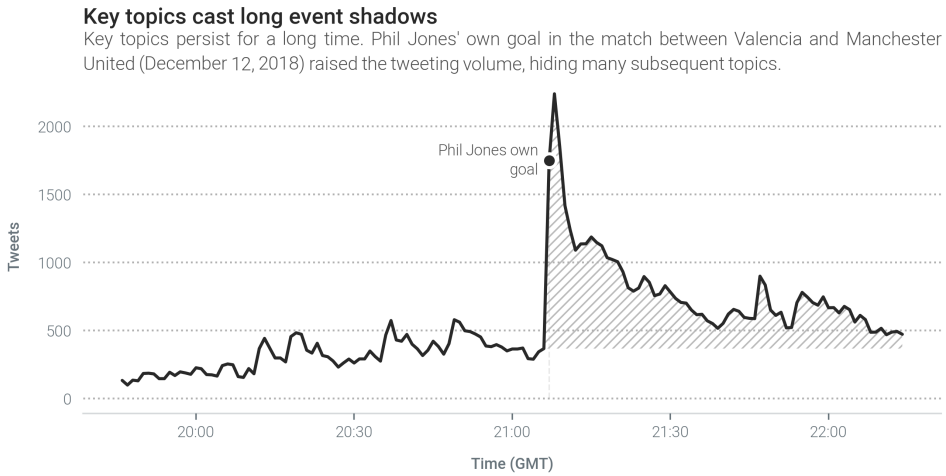

Figure 1: Key topics cast a long event shadow, hiding subsequent topics.

technique validates whether users are discussing topics like breaking news (Mamo and Azzopardi, 2017).

ELD, which we describe next, adapts FIRE's model to process tweets in real-time and adds a novel feature-pivot algorithm to understand topics better.

3 METHODOLOGY

In this section, we describe ELD, a system that we had outlined briefly (Mamo et al., 2019). ELD builds on FIRE's model (Mamo and Azzopardi, 2017), but it operates in real-time and harnesses the link between the document-pivot and feature-pivot techniques. In this paper, we focus on specified, or planned, events, such as football matches. We adopt the following definition of events from Mamo et al. (2019):

Definition 1. *An event is a real-world occurrence that happens at a particular place and at a particular time, and that involves any number of participants.*

In ELD, we argue that machines need to understand events to build better timelines. Therefore we split the event into two parts: an understanding period before the event starts and the actual event period. We describe both periods next.

3.1 Understanding Period

The understanding period is a time for the TDT algorithm to learn about the event before it has started. In this paper, we use the understanding period to ease TF-IDF's difficulties during the event period. TF-IDF's IDF component has difficulty identifying important keywords because tweets are so short. Therefore we replace TF-IDF with Term Frequency-Inverse Corpus Freqeuncy (TF-ICF), a term-weighting scheme that approximates IDF using a reference corpus (Reed et al., 2006). In ELD, this reference corpus is a dataset of tweets collected during the understanding period.

Replacing TF-IDF with TF-ICF gives our TDT approach several benefits. First, TF-ICF fulfils TF-IDF's role of penalizing common words that are not already considered to be stopwords, like *actually*. Second, TF-ICF also penalizes other common terms in the event's vocabulary, such as the names of teams or players in football matches. Consequently, TF-ICF promotes terms that appear during the event, but not before it, such as *goal* or *foul* in football matches.

3.2 Event Period

The event period is the time when the actual event is ongoing. During the event period, ELD receives tweets, pre-processes them, and detects emerging topics using a combination of document-pivot and feature-pivot techniques. We discuss each step next.

3.2.1 Filtering and Pre-processing

ELD takes a very conservative approach to tweet filtering, allowing fine-grained topic detection and facilitating TDT on events with low coverage. Like FIRE (Mamo and Azzopardi, 2017), ELD removes tweets with more than two hashtags. ELD also filters tweets by authors who have never liked a tweet before or who average less than one follower per thousand tweets. In ELD, we added two new filters, removing tweets if they have more than one URL or whose authors have empty profile descriptions.

ELD tokenizes and vectorizes the remaining tweets by removing all stopwords, numbers and

URLs, and splitting hashtags with camel-case notation. To reduce sparsity in the Vector Space Model (VSM), ELD stems all tokens and normalizes repeated characters, replacing words like *goooaaal* with the simpler *goal*. Finally, ELD weights terms using TF-ICF and normalizes the vectors before clustering.

3.2.2 Document-pivot

To determine what users are talking about, ELD uses an on-line clustering algorithm, similar to the one used in FIRE (Mamo and Azzopardi, 2017). As tweets stream in, ELD adds them to the most similar cluster if the cosine similarity exceeds a threshold, empirically set to 0.5. If there are no similar clusters for a tweet, the algorithm creates a new cluster for it. To control performance, ELD freezes clusters that do not absorb any tweets for a period of time, which we call the freeze period. Moreover, ELD discards tweets published more than one time window ago during passionate moments, as are goals in football matches. As a result, ELD can report subsequent topics without undue delays. We describe time windows in more detail in Subsection 3.2.3.

Even at this point, ELD filters clusters cautiously. ELD removes clusters that are made up of replies in their majority, and clusters with a high intra-similarity, calculated as the average similarity between tweets and their cluster's centroid. The latter filter eliminates clusters with minimal variety, such as groups of retweets. By having a document-pivot approach cluster all incoming tweets, ELD can identify key and non-key topics even if they happen in close succession. ELD considers all clusters with as few as three tweets as potential topics, but validates them using a feature-pivot approach.

3.2.3 Feature-pivot

Like FIRE, ELD uses a feature-pivot technique to verify whether a cluster's topic is being discussed like breaking news. We make two changes to FIRE's feature-pivot method. First, we change how ELD compares the popularity of terms over time. FIRE bins tweets into five-minute time windows, which hurts timeliness. A sliding time window would also add overhead and, in turn, hurt timeliness. Instead, we use snapshots, which record the general discourse over time, or the global context. Each snapshot contains the nutrition, or popularity, of all terms observed in the previous 30 seconds:

$$n_{w,s} = \sum_{t \in T_s} t_w \tag{1}$$

The nutrition of word w in snapshot s, $n_{w,s}$, is a summation of the term's TF-ICF weights, t_w, across all tweets in snapshot T_s. We rescale the nutrition values in each snapshot to be between 0 and 1, which also binds burst, described next, between -1 and 1.

Second, we propose a novel feature-pivot technique instead of Cataldi et al. (2013)'s algorithm, used in FIRE. The new algorithm calculates the bursts of terms by comparing their popularity in a cluster, the local context, with their past popularity in all tweets, or the global context. ELD creates the local context similarly to the global context by creating a snapshot from the cluster using Equation 1. Then, ELD computes the burst of a word w at time t, $burst_{w,t}$:

$$burst_{w,t} = \frac{\sum_{g=t-s-1}^{t-1} (n_{w,l} - n_{w,g}) \cdot \frac{1}{\sqrt{e^{t-g}}}}{\sum_{g=1}^{s} \frac{1}{\sqrt{e^{g}}}} \tag{2}$$

The first component measures the change in nutrition of word w between the local context, $n_{w,l}$, and the most recent snapshots, $n_{w,g}$. The second component is a damping factor, which gives more importance to recent snapshots. As a result, ELD only needs to retain the s most recent snapshots since older information has minimal effect on burst. Combined with the earlier re-scaling, the denominator binds burst between -1 and 1. The lower end of a burst indicates that a term is losing popularity, and the higher end indicates that a term is gaining popularity. ELD accepts clusters as topics if they have:

- One term with a burst of 0.8 or higher, or
- Two or more terms with a burst higher than 0.5.

3.2.4 Tracking

Throughout the event, ELD constructs a timeline, or a list of nodes that store topics. This timeline serves two functions: a chronological organization of the topics and a tracking function for long-term topics. Key topics create a phenomenon previously observed by Lanagan and Smeaton (2011): the event shadow. As shown in Figure 1, Twitter users continue discussing key topics for a long time after the fact. The event shadow and clustering's fragmentation call for a robust tracking component, served by ELD's timeline.

To track topics, ELD considers when a topic burst and its similarity with past developments. ELD assumes that topics that burst within 90 seconds of each other, set empirically based on the observed lifetime of topics in volatile domains, belong to the same node. If the most recent node is older than 90 seconds, we compare the new topics with older ones in the previous ten minutes. ELD compares the bursty terms of

Table 1: The datasets used in the evaluation.

Match	Date	Tweets		
		Understanding period	Event period	Total
Manchester United - Arsenal	December 6, 2018	41,586	212,729	254,315
Liverpool - Napoli	December 11, 2018	15,841	165,132	180,973
Valencia - Manchester United	December 12, 2018	3,005	79,419	82,424
Liverpool - Manchester United	December 16, 2018	29,785	303,982	333,767
Tottenham - Wolves	December 29, 2018	3,563	93,223	96,786
Crystal Palace - Chelsea	December 30, 2018	8,937	63,891	72,828

Table 2: The number of key and non-key topics across all datasets.

Match	Goals	Half start and end	Yellow and red cards	Substitutions
Manchester United - Arsenal	6	4	6	6
Liverpool - Napoli	2	4	5	6
Valencia - Manchester United	3	4	4	6
Liverpool - Manchester United	5	4	2	5
Tottenham - Wolves	4	4	5	4
Crystal Palace - Chelsea	3	4	1	5

the new topics with the bursty terms of previous topics using cosine similarity. If the highest similarity with a node exceeds 0.6, also set empirically, ELD adds the new topic to this node. Otherwise, ELD creates a new node for the topic.

In this paper, we consider ELD to have three parameters that control scalability and noise: the minimum cluster size, the freeze period, and the maximum intra-similarity between tweets and their cluster's centroid. We evaluate ELD in the next section.

4 EVALUATION

In this section, we evaluate ELD's ability to create fine-grained timelines of events. In our experiments, we use datasets with different sizes and contrast ELD with Zhao et al. (2011)'s feature-pivot technique. All data and tools used in this evaluation are available in a GitHub repository[1]. The rest of this section describes our evaluation set-up before discussing the results.

4.1 Evaluation Set-up

4.1.1 Datasets

Like many other TDT approaches (Meladianos et al., 2015; Gillani et al., 2017), we focus our evaluation on football matches. These events generate many tweets, their key and non-key topics are easily-enumerable, and they have widely-available ground truth.

We collected our own data for this evaluation because Twitter does not allow full corpora to be shared. Tweet IDs can be provided to download the original tweets, but Weiler et al. (2019) showed that a large number of tweets still cannot be retrieved if they have been deleted, making datasets incomparable. Therefore we used the Tweepy library[2] to collect data from six football matches with varying popularity, as shown in Table 1. We collected English tweets that mention the match's hashtag, or the names of the teams, their coaches and players.

For each football match, we collected two datasets, corresponding to the understanding and event periods. The understanding period starts 75 minutes before the match starts and lasts for one hour, which is when clubs publish their line-ups for the match. Knowing who the participants are keeps TF-ICF from overestimating their importance during the event, which would be detrimental to topical keywords, like *goal* or *half*. The event period starts 15 minutes before the match and lasts for two and a half hours. This period is longer than the matches to ensure full coverage, but we consider only topics that ELD captured while the match was ongoing.

4.1.2 Ground Truth and Evaluation Metrics

In this paper, we seek key and non-key topics: goals (including disallowed goals), the start and end of each half, yellow and red cards, and substitutions. We collected the ground truth for these topics, summarized in Table 2, from LiveScore.com. ELD also captures several topics that are not as easily-enumerable, such

[1] github.com/NicholasMamo/eld-data, last accessed on July 17, 2021

[2] tweepy.org, last accessed on July 25, 2021

Table 3: The configurations of ELD and the baseline for each dataset.

	ELD			Zhao
Match	Cluster size	Freeze period (s)	Max. intra-similarity	Post rate
Manchester United - Arsenal	5	5	0.8	1.5
Liverpool - Napoli	5	10	0.85	1.7
Valencia - Manchester United	3	10	0.85	1.7
Liverpool - Manchester United	7	5	0.85	1.5
Tottenham - Wolves	4	20	0.85	1.7
Crystal Palace - Chelsea	3	20	0.85	1.7

as scoring opportunities. We verified these topics using The Guardian's minute-by-minute reports.

As is common in TDT, we annotated timelines manually using standard IR metrics: precision, recall and the F-measure. We calculated recall only for the enumerable topics listed above. When measuring precision, however, we also accepted other developments of general interest, such as goalscoring opportunities. We rejected spam and other noise, such as opinion-based topics, as well as repeated topics.

4.1.3 Techniques

Unfortunately, few TDT algorithms are made available as open-source projects (Weiler et al., 2019), and FIRE does not operate in real-time. Since we had to implement a baseline ourselves, and even small changes could drastically affect results (Weiler et al., 2016), we implemented a simple feature-pivot technique: Zhao et al. (2011)'s. The algorithm looks for bursts in volume using sliding time windows, which it splits into two halves. If the second half of the window has at least 1.7 times more tweets than the first half, the algorithm accepts it as an emergent topic. The length of the time window starts from 10 seconds and increases to 20, 30 and 60 seconds if the shorter time windows do not have spikes.

In our implementation of Zhao et al. (2011)'s algorithm, we made two changes. First, we modified the post rate depending on the event's volume to optimize results. Second, similarly to ELD, we assumed that bursts that occur within 90 seconds of each other are part of the same topic. Since the baseline does not extract keywords, we generated summaries using Carbonell and Goldstein (1998)'s Maximal Marginal Relevance (MMR) algorithm.

We compare Zhao et al. (2011)'s algorithm with ELD, focusing on granularity over timeliness. Therefore we simulated the event streams at half the speed to minimize the impact of ELD's clustering bottleneck on the timelines. When annotating ELD's timelines, we mainly looked at the topical keywords extracted by our feature-pivot algorithm. Whenever these keywords did not describe the topic adequately, we used summaries created by the MMR algorithm, with the topical keywords serving as the query. Table 3 lists the configurations of the two techniques.

4.2 Discussion

In this subsection, we analyse the differences between ELD and Zhao et al. (2011)'s algorithm, highlighting the limitations of both. The results shown in Table 4 clearly show the difference between the two algorithms. ELD produced significant gains in performance over the baseline; a paired samples t-test shows ELD's improvements to be statistically-significant at the 95% confidence level in precision, and at the 99% confidence level in recall and F-measure.

While ELD's improvements in precision, recall and the F-measure were statistically-significant, the two algorithms detected a comparable number of topics. The contrast seems unexplainable, especially since we present ELD as a fine-grained method. The difference lies in what the two algorithms captured. Zhao et al. (2011)'s algorithm was particularly susceptible to reporting noise and redundant topics. Therefore whereas ELD and the baseline captured a similar number of topics, ELD was more precise.

Moreover, Zhao et al. (2011)'s algorithm suffered with Lanagan and Smeaton (2011)'s event shadows. For example, Phil Jones' own goal in the match between Manchester United and Valencia generated a lasting wave of anger and ridicule, as shown in Figure 1. The tweeting volume remained high for a long time and produced few clear peaks. As a result, the baseline missed several key and non-key topics, including a goal, bringing recall down.

Still, Zhao et al. (2011)'s difficulties were not all due to event shadows. Precision and recall were poor throughout because the algorithm scales poorly. In very popular events, the Twitter API cap of 50 tweets per second eliminates peaks in volume. In unpopular events, the tweeting volume is more volatile, which creates noisy peaks, often with repeated information.

ELD also captured some noise, but much less than the baseline. Like Zhao et al. (2011)'s technique, event shadows also led ELD to capture some late reac-

Table 4: The number of topics captured by ELD and the baseline, and the methods' precision, recall and F-measure scores.

	Topics		Precision		Recall		F1	
Match	ELD	Zhao	ELD	Zhao	ELD	Zhao	ELD	Zhao
Manchester United - Arsenal	35	24	0.829	0.625	0.636	0.409	0.720	0.494
Liverpool - Napoli	43	20	0.744	0.600	0.706	0.176	0.725	0.272
Valencia - Manchester United	31	32	0.581	0.531	0.588	0.353	0.584	0.424
Liverpool - Manchester United	42	10	0.714	0.800	0.813	0.250	0.760	0.381
Tottenham - Wolves	27	52	0.593	0.423	0.647	0.588	0.619	0.492
Crystal Palace - Chelsea	37	38	0.703	0.474	1.000	0.615	0.826	0.535
Macro-average	36	29	0.694	0.576	0.732	0.399	0.706	0.433

Table 5: A breakdown of recall results from the ground truth.

Type	ELD	Zhao
Goals	1.000	0.870
Half start and end	0.667	0.333
Yellow and red cards	0.522	0.261
Substitutions	0.688	0.188

tions to key topics. Our algorithm also captured noise, such as opinions and spam. Still, noise did not overwhelm ELD's timelines, especially when considering that our method accepted clusters with as few as three tweets. However, ELD would still benefit from better topic tracking and more robust filtering to minimize subjectivity. More importantly, neither the noise nor event shadows kept ELD from capturing non-key topics, as shown in Table 5. Our approach captured all goals, and more than twice as many non-key topics as Zhao et al. (2011)'s algorithm. In addition, ELD could identify other interesting topics, such as injuries and goalscoring opportunities.

Nevertheless, the breakdown of results shows ELD to struggle more with non-key topics than with key topics, raising an important question: what makes ELD different from traditional TDT approaches if it inherits their difficulties? A closer look at the missed non-key topics shows that ELD overcame the technical difficulties of traditional methods, but not the behavioural challenges of Twitter. In other words, ELD missed topics because few Twitter users were discussing them, not because of the event shadow. For example, in the noisy aftermath of Phil Jones' own goal, ELD still captured two substitutions and the last goal, which the baseline missed.

We identified two challenging situations for ELD: bias and particular scenarios. Bias can manifest itself between popular and unpopular teams, like in the match between Tottenham and Wolves. However, bias also appears between English and non-English teams, such as between Manchester United and Valencia. In either case, ELD had difficulty identifying non-key topics about the less popular team, which generates fewer tweets, or fewer tweets in English. In fact, ELD captured all of Manchester United's substitutions, but none of Valencia's.

The particular scenarios in individual matches also influenced tweeting behaviour. ELD missed two Liverpool yellow cards towards the end of their match against Napoli, even though the dataset's imbalance favoured Liverpool. Since the referee gave these yellow cards so late, they had little effect on the match. As a a result, relatively few Twitter users commented about the yellow cards, leading ELD to miss both.

ELD's consistent results throughout all events further prove that ELD overcame the technical limitations. In fact, ELD's best recall results came in the match between Crystal Palace and Chelsea—the smallest dataset. This match had a low tweeting volume, but no notable event shadows or unexpected noise. In contrast, exaggerated swings in volumes, often caused by key topics, pose a greater difficulty to ELD. In the match between Manchester United and Valencia, the tweeting volume went from a consistently low baseline of tweets to a much higher volume after the own goal. As a result, ELD's parameters would either miss many topics in the first phase and reduce noise in the second one, or the opposite.

Aside from these challenges, ELD shows the strength of combining document-pivot and feature-pivot approaches. ELD marks a significant improvement over the baseline by overcoming the individual technical limitations of document-pivot and feature-pivot approaches. Nevertheless, TDT needs to handle the behavioural challenges too. We conclude with suggestions for future work next.

5 CONCLUSION

Twitter gave a lot to TDT, but real-time access to large volumes of tweets is meaningless if TDT does not overcome the social network's unreliability. However, unreliability only hides newsworthy content, not eliminate it. With ELD we showed that machines can identify the news among this noise. ELD over-

came the technical limitations of document-pivot and feature-pivot approaches, and recalled more key and non-key topics than Zhao et al. (2011)'s algorithm. Beyond the gains in performance, the combination of document-pivot and feature-pivot techniques allowed ELD to understand events better by extracting both tweets and topical keywords.

Still, TDT's road to harness Twitter's advantages remains a long one. It is still necessary to find a balance between detecting non-key topics during an event and minimizing noise, especially in emotionally-charged events. Moreover, while eliminating noise remains a significant challenge, tracking also needs to improve to minimize redundant reporting, which remains an understated problem in TDT.

ACKNOWLEDGEMENTS

The research work disclosed in this publication is partially funded by the Endeavour Scholarship Scheme (Malta). Scholarships are part-financed by the European Union - European Social Fund (ESF) - Operational Programme II - Cohesion Policy 2014-2020 *"Investing in human capital to create more opportunities and promote the well-being of society"*.

REFERENCES

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

Carbonell, J. and Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 335–336, Melbourne, Australia. Association for Computing Machinery.

Cataldi, M., Caro, L. D., and Schifanella, C. (2013). Personalized Emerging Topic Detection Based on a Term Aging Model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):1–27.

Farzindar, A. and Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31(1):132–164.

Gillani, M., Ilyas, M. U., Saleh, S., Alowibdi, J. S., Aljohani, N., and Alotaibi, F. S. (2017). Post Summarization of Microblogs of Sporting Events. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 59–68, Perth, Australia. International World Wide Web Conferences Steering Committee.

Ifrim, G., Shi, B., and Brigadir, I. (2014). Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering. In *Proceedings of the SNOW 2014 Data Challenge*, page 33–40, Seoul, Korea. CEUR.

Lanagan, J. and Smeaton, A. F. (2011). Using Twitter to Detect and Tag Important Events in Live Sports. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, page 542–545, Barcelona, Spain. Association for the Advancement of Artificial Intelligence.

Mamo, N. and Azzopardi, J. (2017). FIRE: Finding Important News REports. In Szymański, J. and Velegrakis, Y., editors, *Semantic Keyword-Based Search on Structured Data Sources*, pages 20–31, Gdansk, Poland. Springer International Publishing.

Mamo, N., Azzopardi, J., and Layfield, C. (2019). ELD: Event TimeLine Detection – A Participant-Based Approach to Tracking Events. In *HT '19: Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 267–268, Hof, Germany. ACM.

Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., and Vazirgiannis, M. (2015). Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, page 248–257, Oxford, United Kingdom. The AAAI Press.

Reed, J. W., Jiao, Y., Potok, T. E., Klump, B. A., Elmore, M. T., and Hurson, A. R. (2006). TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams. In *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*, pages 258–263, Orlando, Florida, USA. IEEE.

Saeed, Z., Abbasi, R. A., Maqbool, O., Sadaf, A., Razzak, I., Daud, A., Aljohani, N. R., and Xu, G. (2019a). What's Happening Around the World? A Survey and Framework on Event Detection Techniques on Twitter. *Journal of Grid Computing*, 17(2):279–312.

Saeed, Z., Abbasi, R. A., Razzak, I., Maqbool, O., Sadaf, A., and Xu, G. (2019b). Enhanced Heartbeat Graph for Emerging Event Detection on Twitter using Time Series Networks. *Expert Systems with Applications*, 136:115–132.

Unankard, S., Li, X., and Sharaf, M. A. (2015). Emerging Event Detection in Social Networks with Location Sensitivity. *World Wide Web*, 18:1393–1417.

Weiler, A., Beel, J., Gipp, B., and Grossniklaus, M. (Oct 13, 2016). Stability Evaluation of Event Detection Techniques for Twitter. In *Lecture Notes in Computer Science book series (LNCS, volume 9897)*, page 368–380, Stockholm, Sweden. Springer.

Weiler, A., Schilling, H., Kircher, L., and Grossniklaus, M. (Jun 14, 2019). Towards Reproducible Research of Event Detection Techniques for Twitter . In *2019 6th Swiss Conference on Data Science (SDS)*, pages 69–74, Bern, Switzerland. IEEE.

Zhao, S., Zhong, L., Wickramasuriya, J., and Vasudevan, V. (2011). Human as real-time sensors of social and physical events: A case study of twitter and sports games.