Video Retrieval for Multimedia Verification of Breaking News on Social Networks

Lyndon J B Nixon, Shu Zhu, Fabian Fischer
MODUL Technology
Am Kahlenberg 1
Vienna, Austria
nixon@modultech.eu

Walter Rafelsberger, Max Göbel, Arno Scharl
webLyzard technology
Puechlgasse 2/44
Vienna, Austria
scharl@weblyzard.com

ABSTRACT

This paper presents an approach to automatically detecting breaking news events from social media streams, using event detection to collect in near real time relevant video documents from social networks regarding that breaking news. A visual analytics dashboard provides access to the results of the content processing pipeline, providing a rich interactive interface to explore emerging stories and select video material around those stories for verification.

CCS CONCEPTS

•Information systems →Clustering; Search interfaces; Clustering and classification; Retrieval effectiveness; *Query representation; Presentation of retrieval results;*

KEYWORDS

Video Retrieval, Social Network Retrieval, Social Media Retrieval, Social Media Extraction, Breaking News Detection, Story Detection

ACM Reference format:
Lyndon J B Nixon, Shu Zhu, Fabian Fischer and Walter Rafelsberger, Max Göbel, Arno Scharl. 2017. Video Retrieval for Multimedia Verification of Breaking News on Social Networks. In *Proceedings of MuVer'17, October 27, 2017, Mountain View, CA, USA.*, , 9 pages.
DOI: https://doi.org/10.1145/3132384.3132386

1 INTRODUCTION

Today, when something newsworthy happens somewhere in the world, social media users often capture the event on video and publish their recording on Twitter, YouTube or other social networking platforms. However, alongside genuine user-generated content about newsworthy events almost in almost real time, the resulting journalistic and social interest in such content has encouraged many users to post fake or misleading video material, claiming to be likewise coming from the breaking news event. As a result, it is no longer sufficient to provide journalists with a tool to find (video) content around a breaking news story, but also to aid them in determining whether a video is to be trusted or not. When reasonable doubt exists, such a tool should enable them to additionally verify the video for its authenticity and veracity.

This paper addresses this requirement by presenting work done in the InVID project (In Video Veritas; www.invid-project.eu) to detect breaking news stories and retrieve relevant video from social networks for those stories as a first step before deciding to engage in fuller video verification. Firstly, we will look at related work in the different areas of breaking news detection and news video retrieval. We combine both approaches into a holistic workflow, which subsequently feeds into the multimedia verification process. Then we will present InVID work done in each area, focusing on how the platform handles distinct challenges arising from this hybrid approach, and how breaking news detection can support the retrieval of relevant video content from social networks. This leads us into the issue of how to describe breaking news in an appropriate model for information retrieval. Finally, we conclude with an outlook for further improving InVID technologies for breaking news detection and news video retrieval.

2 RELATED WORK

2.1 Breaking News Detection

Recent work on breaking news detection is centered around Twitter as the major source of news stream data. [9] confirmed "Twitters rising potential in news reporting" that can "fundamentally change the way we produce, spread, and consume news". Their analysis showed that "the people who broke the news were able to convince many Twitter users before confirmation came from mass media" using Bin Ladens death as a case study. Thus, it provides an evident motivation for the real-time breaking news detection from the Twitter stream.

Topic modeling is a common approach that can be applied to detect breaking news on Twitter [1, 4, 21]. Topic detection (modeling) algorithms, such as Latent Semantic Analysis (LSA) [5, 11] or Latent Dirichlet Allocation (LDA) [2], provide means to organize a collection of electronic documents into semantically coherent groups (topics) based on the frequent word co-occurrence matrix (e.g. TF-IDF metrics). Topic detection approaches often involve topic clustering, ranking and labeling stages [7, 10, 12, 13, 17, 20]. One common method to find novel (emerging or recent trending) topics from a data stream is looking for bursts in frequent occurrences of keywords and phrases (n-grams) [1, 4, 8, 12, 13]. Even more recently and novel is the use of neural networks. [3] demonstrated encouraging results using the word2vec Skip-gram model to generate event timelines from tweets. [14] achieved an improvement over the state-of-the-art ffirst story detection (FSD) results by expanding the tweets with their semantically related terms using

MuVer'17, October 27, 2017, Mountain View, CA, USA.

ISBN 978-1-4503-5510-0/17/10…$15.00
DOI: https://doi.org/10.1145/3132384.3132386

word2vec. We could also demonstrate in an experimental evaluation that an approach based on character embeddings significantly outperforms the current state-of-the-art in tweet clustering for breaking news detection [19].

The SNOW 2014 Data Challenge [16] has confirmed newsworthy topic detection to be still a challenging task: the top F-score of the competing solutions was only 0.4 (Precision: 0.56, Recall: 0.36). The limitations of current state-of-the-art approaches include early topic detection, topic relevance, topic representation and evaluating the performance of topic detection methods. We will present the InVID approach, which proves to provide a fair balance between real-time performance and story output relevance.

2.2 News Video Retrieval

Most research in video retrieval focuses on the way in which descriptive video metadata, extractable low level features and/or high level semantic annotations may be used to improve precision and recall in the retrieval system. Such research tends to assume the user can control the data that is available to them for the video collection, i.e. that they may prepare data in advance of retrieval e.g. through additional metadata extraction. In social media retrieval, the video collection is too vast and visibility of the items heavily restricted (e.g. Twitters Streaming API only provides less than 1% of live tweets). "Social multimedia" indeed presents new opportunities alongside new research challenges [15]. In social multimedia research, we observe that a finite video collection is created from the social network data based on some sampling technique (generally based on filtering the document space according to the use case, e.g. through geolocation or significant term match) and then feature-based or semantic retrieval is tested on this finite collection. This is different from our goal, which is to maximise the relevance of the video documents returned for our dynamically created searches based on breaking news detection, prior to any post-processing (for verification).

Since document retrieval is via the social network APIs, search result quality can only be controlled by us through the query formulation. Search APIs primarily take text based queries and match against descriptive metadata of the documents, i.e. the user-provided title and description. Additional parameters may be supported, e.g. filter by time uploaded or natural language. We have analysed the query expressiveness possible for each social network (Table 1). Research in query construction considers the means to construct the query from another accessible data source, e.g. online news articles [18], and how it affects quality of retrieval. Our source for queries is the breaking news detection, i.e. our retrieval quality is directly related to how our breaking news output is modelled. Query construction can include other Information Retrieval techniques e.g. to overcome natural language limitations by query expansion [6]. We focus our queries by requesting only recently uploaded documents, since breaking news video can not be posted before the breaking news occurs.

Platform	Endpoint	Query	Temporal	Geo	Lang
YouTube	search	string	+	+	+
DailyMotion	data	string, tags	+	+	+
Vimeo	search	string	-	-	-

Table 1: Video API comparison

3 BREAKING NEWS DETECTION

In InVID, we have the goal to automatically identify newsworthy events which could guide journalists to online media being posted in association with that event (and which may require verification before it can be used in the professional news cycle). We had three primary requirements to address in the InVID context, which took our work away from the classical research activities in this area:

- Timeliness of detection of a new newsworthy event;
- Addressing multilinguality and alternative names in the detection approach;
- Quantifying the newsworthiness of the event as suitable for extracting eyewitness media.

We established that social media streams, in particular Twitter, are the most effective sources of data for this task and developed a workflow model for story detection (Fig. 1) which can be explained in terms of the current implementation thus:

- Content modeling - we model each tweet as a bag of keywords based on natural language processing and perform keyword alignment based on named entity recognition;
- Clustering - we chose a community detection algorithm as a means to cluster tweets and configured our approach to better disambiguate between distinct stories and merge overlapping stories;
- Burst detection - we experimented with organizing stories based on burst detection approaches to highlight recent events;
- Ranking - we provide alternative ranking possibilities(e.g. volume, frequency) and classify stories according to predefined topics (based on the IPTC NewsCodes).

We base story detection on content extracted from the Twitter Streaming API, using a manually set-up list of 61 professional news accounts from around the world. We have implemented a set of pre-processing measures to remove irrelevant data from the stream: a word blacklist to strip out spam, a length filter (of 30 characters) to ensure the tweet has enough textual content, as well as a language check as part of the NLP pipeline to ensure the text is in a supported language (currently: English, French and German). To cluster each tweet into distinct stories, appropriate models are needed that reduce the complexity and support the task of clustering, which is based on computational calculation of similarity between documents. Given the scale of documents to be clustered, the model should also support a computationally inexpensive clustering method. The Baeza-Yates algorithm for indexing large text corpora by statistically significant n-grams has been provably efficiently scalable for approximate string matching tasks. Using this algorithm, we model each tweet as a set of n-grams as keywords, with tf-idf measures avoiding that overly frequent or extremely rare strings become keywords. With each tweet modelled as a bag of (key)words, we have settled on the Louvain Modularity algorithm for clustering. The algorithm is more commonly used to detect communities within social networks - in our case we detect "communities" of related keywords over time. We chose this algorithm

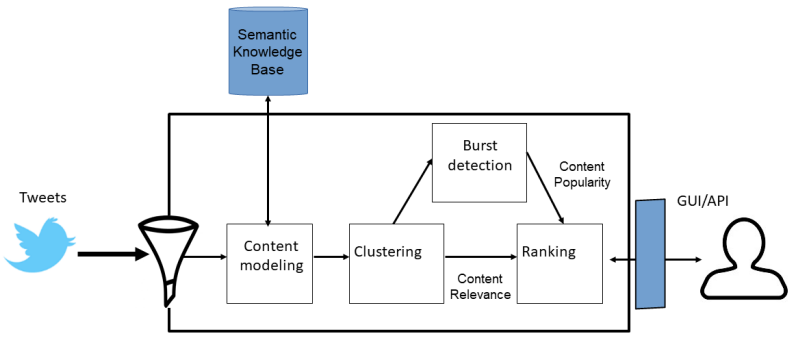

Figure 1: InVID workflow for story detection.

as the graph of keywords is specifically structured in the same way as a community of social network contacts and it proved to be much more efficient than k-means as well as performing more scalably than an approach calculating character embeddings with tweet2vec which we also experimented with. For the latter, we would need to be able to update the training model for the neural network more rapidly than the time needed for the new model could be created. The output is a set of document clusters based on the keyword co-occurrences, and we label each cluster with the top three keywords by co-occurrence weight. Through observation of the results, we clean up some spurious clustering by pre-processing keywords (aligning partially overlapping n-grams) and testing different thresholds for the graph partition algorithm. We rank the clusters by size (number of documents as members of the cluster). Fig. 2 shows the top three stories on 14 June 2017, dominated by the London apartment tower fire.

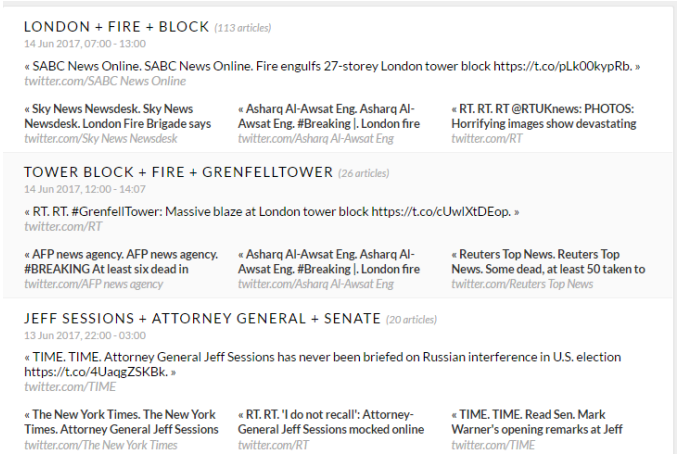

Figure 2: Top 3 stories on 14 June 2017.

To evaluate the effectiveness of our breaking news detection, we need to agree first on the methodology to be followed. The dataset used in prior evaluations (SNOW 2014) is not used as we want to test on real data being fed into the InVID platform and the methodology of the SNOW Challenge is different from our set-up, using 5 time slots of 15 minute duration as the basis for comparison, where we look at news detected over a 24 hour period. This of course raises the question of how we will get our ground truth and a measure for how to compare the results of our algorithm with the ground truth.

SNOW 2014 did confirm newsworthy story detection to be a challenging task: F-score: 0.4, Precision: 0.56, Recall: 0.36 [10]. We choose to evaluate two metrics of our story detection: quality and correctness. Correctness of the stories can be measured as a factor of whether a cluster can be unambiguously identified as representing a distinct story (newsworthy action or event). Hence the observer must not decide on the newsworthiness of the story, but simply that the cluster can be seen as representing a potential story. While this needs to be reflected by the (top) documents in the cluster, since we will look at cluster quality below, we will use the story label in this case as the determinant of the story of the cluster. Since different labels (which determines the separation between clusters) may, in fact, be referring to the same news story we add a complementary measure for distinctiveness, i.e. how many of the clusters identifiable as stories actually represent distinct news stories. Clustering quality can be measured in terms of completeness and homogeneity. Two structural observations can be made on a list of clusters: whether a cluster actually merges two or more stories or whether a story is represented by two or more different clusters. This requires that every document in a cluster is individually marked as belonging in its cluster, in a different cluster, or in no cluster. Completeness refers to the extent to which documents of a story are in the same cluster. Homogeneity refers to the extent to which clusters only contain documents of one particular story.

We will conduct a ”live” evaluation, i.e. in real time using the story results visible in the InVID Dashboard, a Web based portal

interface to the data collected by the InVID Platform. It will be a manual exercise, as the human evaluator will make the assessment about the quality of the stories. Values will be drawn through a daily assessment of the stories detected from our Twitter Accounts stream used for breaking news detection, as the Dashboard allows us to see and browse the top ten stories detected in the last 24 hours.

The lead author conducted the story evaluation over the period June 19 to June 23, 2017. For each story, we evaluated the label (does it meaningfully refer to a news story? does it reflect a single story or multiple stories?) and the documents presented for that story (do they relate to the story represented by its label?). From this, we could combine our insights into 4 metrics which we can compare across sources and days:

- Correctness: are the generated clusters correctly related to newsworthy stories?
- Distinctiveness: does each individual cluster precisely relate to an individual story?
- Homogeneity: are the documents in the cluster only relevant to the newsworthy stories represented by the cluster?
- Completeness: are the documents in the cluster relevant to a single, distinct news story?

Fig. 3 illustrates the daily evaluation: this shows the top ten stories in the dashboard on the first day, June 19 2017. All ten are news, although some represent a merge of two distinct stories (the stories are numbered and some have two numbers in the "distinct" column). For example PORTUGAL + FIRE + AFGHANISTAN clearly referred to both fires in Portugal and in Afghanistan. For the documents provided for the story (Nr Docs column), most distinct stories also have relevant documents (This Story column). We had here one story US NAVY + USS + JAPAN where some irrelevant documents were associated with it. Non-distinct stories typically had documents associated with them from both distinct stories (This Story and Other Story columns). TRUMP + INVESTIGATION + CUBA only had documents from one story although the label refers to two (Trump and the Russia investigation, Trump and Cuba policy), note how this affects the Homogeneity score which provides the ratio of relevant documents to all documents for all stories represented by the label (it does not penalise distinctness). In comparison, Completeness does penalise merged stories as in the case of PORTUGAL + FIRE + AFGHANISTAN since only the documents from the primary story are counted as relevant out of the set of all documents for that story. This analysis was repeated in the same way for the following four days.

The average of the values over all stories for the five days is:

- Correctness: 0.895 (The 5 non-stories out of the total 46 were all generated from the tweets of one news organisation, we found it was tweeting very often and many tweets were commentary rather than news stories so we will remove this account for the next phase of story detection.)
- Distinctiveness: 0.598 (Values varied from 0.44 to 0.71 showing this is still an issue of concern. It penalises both merged stories and split stories, both of which occurred daily.)
- Homogeneity: 0.93 (It is confirmed that the documents are nearly always relevant to the story. The few irrelevant documents that were found were always related to a story label being a merge of two stories, but documents from one of those stories were missing or a story label related to one story, but documents from another story were present.)
- Completeness: 0.93 (While the values varied slightly with homogeneity, the average is the same. This highlights that most clusters were indeed single stories with relevant documents. In two of the five days, completeness scored lower than homogeneity due to single clusters referring to two stories and containing documents from both, since completeness penalizes this.)

The current approach already approaches 100% for clustering documents into newsworthy stories (it seems, if we removed the one Twitter account that generated the non-stories, we would have scored 100% here), which reflects the quality of the data source chosen for the task - Twitter accounts of professional news channels. We compared this with the correctness score achieved from a Twitter stream which takes user tweets which mention "breaking news". Here, we averaged a correctness of 0.76 which reflects the expected lower data quality of user tweets. Distinctiveness is consistently lower. While there are merged stories, the most common issue is split stories, i.e. stories with a higher volume of tweet reporting tend to diverge more with respect to the words used to report them which in turn leads to a larger set of distinct keywords being extracted. As merging clusters is based on the overlap between the keyword sets that define them, there is still a point where the keyword variation is too high to affect a merge. Finally, homogeneity and completeness scores are consistently high across the stories and across all five days.

Thus, our own evaluation evaluation shows that the breaking news story detection performs very well, and for the stories we find, the clustering is also very accurate for ordering documents (tweets) to the correct story.

4 NEWS VIDEO RETRIEVAL

Based on the story detection, we need to address the most appropriate process to select the relevant video content from social networks for those stories. We set up a social media extraction pipeline which is configurable and extendible to support additional sources. The initial pipeline supports YouTube, DailyMotion and Vimeo APIs. A first experiment compared the quality of results from YouTube when querying with single terms (e.g. Obama), term combinations (e.g. Obama+dinner) or multiple terms (e.g. Obama White House dinner). We found term combinations worked best to return relevant documents (when the query is applied for the time period in which the news event the terms refer to was currently relevant). Single terms produced too many false positives while multiple terms too many false negatives (inferrable by the observation that term combinations returned relevant documents the multiple terms query did not).

At the outset, the only data we had to use from the Twitter news stream in the platform was an aggregation of keywords extracted from the tweets and their top associations (keywords which co-occur with them). We took the most frequent 10 keywords and for each keyword the top 10 associations, thus generating 100 keyword pairs which are then used to perform conjunctive queries over each API at a 6 hourly interval, filtering out query duplicates and

STORIES	Label	Story Y/N	Distinct	Nr Docs	This story	Other story	No story	Homog.	Compl.
#1	LONDON VAN MOSQUE	Y	1	10	10			1	1
#2	MALI ATTACK RESORT	Y	2	10	10			1	1
#3	FIRE DEAD GRENFELL TOWER	Y	3	10	10			1	1
#4	PORTUGAL FIRE AFGHANISTAN	Y	4.5	10	3	7		1	0.3
#5	LONDON MOSQUE TERRORIST ATTACK	Y	1	10	10			1	1
#6	US NAVY USS JAPAN	Y	6	9	6		3	0.666667	0.666666667
#7	TRUMP INVESTIGATION CUBA	Y	7.8	5	5			0.5	1
#8	PORTUGAL DEAD COLLISION	Y	4.6	5	2	3		1	0.4
#9	ISLAMIC STATE IRAN SYRIA	Y	9	3	3			1	1
#10	JET PM LEE RUSSIAN	Y	10.11	1	1			0.5	1

Figure 3: Story detection evaluation for 19 June 2017

applying a time restriction to video posted within 24 hours of the query. With this, we collected 700-1100 news videos daily with the social media extraction pipeline.

When singular news events dominate the global news coverage, our stories tended to be dominated by clusters using multiple different keywords associated with the same dominant story (this case of 'split stories" was noted in the previous section). This meant that, while many more stories were being detected from the Twitter stream, video retrieval was tending to be based on queries related to a more narrow set of more dominant stories. With the implementation of a more sophisticated method for story detection and labeling, as an alternative to the keyword-association pairs, we could use the labels of the detected stories. These currently consist of the three most frequent keywords in the cluster representing a story. While a triple term query may reasonably retrieve fewer matching documents than a term pair, the actual volume and relevance of retrieved documents will depend on the quality of the story labelling, just as the current approach depends on the quality of the keywords and their associations.

To comparatively evaluate the two approaches, we can not use classical recall measures since we can not say what is the total number of relevant documents on any social media platform at any time for one query. Precision can be calculated based on the set of results retrieved for each query, where the maximum number of results evaluated will be capped at 20 documents (which is also the first page of video results on YouTube). On the other hand, we should consider whether success in information retrieval only occurs if and only if the retrieved video is relevant to the story represented by the query, or if any newsworthy video being retrieved can be considered a metric for success. Indeed, whereas the timeliness of the story detection is important to ensure stories are detected as they emerge and queries are made for relevant video at the moment the news story is still newsworthy, in the video retrieval videos will continue to be posted for a news story for a longer time after the news story initially occurred and those retrieved videos can still be relevant for discovery and verification in the InVID context. Thus, queries which reference keywords that persist in the news discussion (such as Donald Trump) are likely to return other videos which are not relevant to the current story but still reference an earlier newsworthy story. Since while our precision measure can indicate how many videos in our results are relevant to the query itself, low precision may hide the fact we still collect a high proportion of newsworthy video content. Still, precision may act as an evaluation of the quality of our query to collect media for the specific story. On the other hand, we choose a second precision measure, which we will call "accuracy", which measures the proportion of all newsworthy video returned for a query. Since this measure will include video not directly related to the story being queried for, this acts as an evaluation of the appropriateness of our query to collect newsworthy media generally. Finally, we will include a recall measure which is defined as the proportion of newsworthy video retrieved which is relevant to the story being queried for, ergo our recall is the precision divided by the accuracy and acts as a measure of the specificity of our query for the news story. For comparative evaluation, we would of course prioritize higher precision for our queries, but also since social media retrieval will invariably involve the possibility of false positives (due to the varying relevance of user provided titles and descriptions which determine the results of a query), we will also welcome higher accuracy (once we accept the existence of false positives in the results, we desire to minimize the extent of completely irrelevant video that may be collected as a result).

None of these measures indicates if we achieve a further goal, which is to collect a broader range of video material related to the news. To do this, we need to annotate each query with the story it is querying for. As there is no official classification of news stories each day with which we could annotate queries, we will take a simpler approach. The first query will be annotated as belonging to a story S1. The subsequent query, if we determine it to be querying on the same story, will also be annotated with S1. If it is querying on a different story, we will annotate it with a new story S2. Ongoing, every new story that is being queried will receive a new identifier, following a standard numbering order. As a result, we may conclude with 2 further measures. Story breadth is the sum of unique stories queried for within the set of queries being evaluated. Story depth is the measure of the extent retrieved relevant videos are distributed across distinct stories, i.e. we want to reward a higher average precision value across different stories as opposed to having higher precision for only some stories while other stories suffer from low precision in video retrieval. Since this is a factor of uniform distribution of the precision across stories, we calculate story depth as the mean of the precision of video retrieval across stories (which is itself the average of the precision of video retrieval across all queries related to a story) multiplied by one minus the standard deviation (of the values for average precision

for each story). Hence a more uniform distribution for precision of retrieval across stories will tend towards a value closer to the overall average precision of the queries as a whole (where standard deviation is zero, story depth = overall average precision). However, as the distribution becomes more non-uniform (some stories have higher precision and some have lower precision, so the relevant results start to become skewed towards having more content for a smaller number of stories than what was queried for), standard deviation will increase and the story depth will drop.

For manual evaluation, we will look at two one day's results from the Twitter news accounts stream. We can examine both the current query constructions (keyword-association pairs) and the potential query constructions (story labels) in the dashboard, manually making the queries on the YouTube API to check the list of videos returned. We choose two separate days in order to attempt to consider one day where news coverage has been quite generally spread (we choose the results from the day of writing which is 12/13 June 2017, as there is no single dominating story in the aggregated news) and one day where news coverage has been skewed towards one larger news story (we choose 9/10 May 2017, which covers the firing of FBI director James Comey). We choose a number of 25 queries for the comparison of each approach: for the current approach we take the top-5 keywords and their top-5 associations over all documents; for the potential approach we take the top-5 stories over all documents and the top-5 stories from each of the top four news topics (we classify documents according to the IPTC NewsCodes and order this classification as a list of topics by size).

Metric	10-May	current	proposed
avg precision		0.89	0.52
avg accuracy		0.97	0.69
avg recall		0.91	0.59
f-score		0.92	0.57
story breadth		2	15
story depth		0.85	0.35
Metric	**13-Jun**	**current**	**proposed**
avg precision		0.36	0.54
avg accuracy		0.84	0.82
avg recall		0.42	0.64
f-score		0.425	0.59
story breadth		9	18
story depth		0.17	0.3

Figure 4: Video retrieval results

Looking at the results from a news day where there was a mix of stories in the news reporting (13 June 2017), we observe that the proposed approach achieves a wider breadth of news stories, querying for 18 distinct stories within the top-25 queries whereas the current approach queried for 9 distinct stories. More commonly tweeted stories tend to generate more individual queries in the current approach, e.g. the story on the tornado warning in different US counties had 5 different queries in the current approach while was the subject of 1 specific query in the proposed approach. The average precision for the current approaches' 5 different queries was 0.16 compared to a precision of 0.1 for the specific query in the proposed approach, suggesting one does not retrieve significantly fewer documents for the story with less queries while of course achieving a greater breadth of news story coverage in the total retrieved document set. In fact, while average precision was higher for the proposed approach average accuracy was almost the same, suggesting the same quantity of newsworthy documents may be retrieved by the proposed approach with a higher proportion of them being relevant to the current news stories (recall of the proposed approach was 0.64 compared to 0.42 for the current approach). Similarly, the F-Score for the proposed approach was 0.59 compared to 0.425 for the current approach. The remaining issue for the proposed approach, with its greater breadth of news coverage in the retrieved document set, would be that having less distinct stories in the set could mean having better coverage of those stories in the documents. Our story depth measure for how many specifically relevant documents are retrieved for each of the stories suggests that this is not an issue, as while double as many distinct news stories are covered in the retrieved documents the story depth is 0.30 compared to a value for the current approach of 0.17.

By considering a second news day we can check if the two approaches compare equally when the news context is different. We have observed through the dashboard that the global news coverage being collected in the Twitter Accounts stream does mean that days in which a major news story occurs (with global significance) lead to that story being referred to in the twitter stream at a significantly higher frequency than any other story (which we have termed the "dominant" story). This makes sense: whereas stories of regional interest will only be reported by a subset of our Twitter news accounts, global stories will be covered by potentially all of those accounts leading to a significant difference in volume of documents. Since such "dominant news story days" are an unavoidable aspect of daily news detection, we took the day of James Comey being fired as FBI director by President Donald Trump (10 May 2017). As expected, our keyword based queries almost all relate to this story, since there are multiple top keywords for the same story (trump, fbi, james, house, director, chief). Indeed, 24 of the top-25 keyword pairs for queries were about this story, only "president + korea" referring to another story on that day, the election of Moon Jae-in as the new president of South Korea. The dominant story also generates a lot of video content on the video platforms on that day, as a response to global discussion and reaction to the event, so these queries also show a high average precision. While our story breadth is just 2, our average precision is 0.89. Accuracy is only slightly higher but almost perfect at 0.97; with most retrieved documents relating specifically to the dominant story recall is at 0.91. While there is only one other story to retrieve documents for in the top-25 queries, and that is due to 1 query, it's precision was also high (0.85) and thus story depth (the extent to which all stories in the query list are represented by retrieving relevant documents) stands at 0.85. These high values seem to communicate that this has been a highly effective approach to news video collection but again we must remind ourselves that only 2 different stories are present in that collection, and 96% of the collected and relevant video is about a single story.

In the proposed approach for this news day, there is a significant difference in story breadth. Now 15 stories are distinctly queried

for in the top-25 queries. The lead story is the same dominant story as in the current approach, FBI director James Comey being fired. Three queries are made for this story with an average precision of 0.82 (due to the third story having an irrelevant keyword in its label, otherwise precision was at 1 for the queries based on story labels for this story). Compared to the current approaches precision of 0.89 for this story, this is not much less. Potentially much more video could be collected for the story in the current approach (retrieved by 24 different queries as opposed to 3) - however we have not considered here how much duplication of content occurs in subsequent queries targeting the same story. The number of unique video documents retrieved by both approaches may not differ as much. Overall average precision is 0.52, reflecting that some story queries are very effective and some are not (e.g. "ATTACK + COURT + BRISBANE" which references correctly a news story relevant at that time - a man standing trial in court for an attack in Brisbane, Australia - but failed to return any relevant video documents). Accuracy averages at 0.69 and our recall is 0.59. Finally, story depth is measured as 0.35. The figures are all lower than the current approach but we need to ask if we prefer 85% story depth for 2 stories on the day or 35% story depth for 15 stories, ensuring a much broader choice of video content for news on the platform. Comparing the 10 May proposed approach results with the current approach when it is also querying for a broader range of stories (the 13 June data), we can note that the proposed approach has higher precision (0.52 to 0.36), recall (0.59 to 0.42) and story depth (0.35 to 0.17). This suggests that the proposed approach provides more accurate queries for the stories and hence performs better on retrieving relevant documents across all stories.

We can also consider how the two approaches performed comparatively on both days. It is clear that the current approach is strongly influenced by dominant stories, significantly reducing the story breadth in the collected documents on such days. The proposed approach performed, on the other hand, very similarly on both days despite the clear difference in the distribution of tweets about news stories. It is possible that as story breadth reduces, the current approach performs better on the evaluation metrics (precision, accuracy, story depth) but indeed at the cost that the collected documents cover a smaller number of news stories. However, the proposed approach collects documents from a broader range of stories, and while precision and story depth may then be lower they perform better than when the current approach should be equally broader in its collection and appear that they should be more stable over time regardless of how tweets about news stories are distributed day by day; a consistent precision of around 0.5 and story depth of around 0.3 would indicate that the document collection in the proposed approach would be much more balanced across all the stories.

5 APPLICATION: THE INVID DASHBOARD

To enable users to explore the detected stories and the collected media around those stories, we have integrated the data results into a front end user interface called the InVID dashboard. Stories and documents from the supported Video sources (YouTube, Vimeo, DailyMotion and - after the above evaluation was completed - Twitter Video) are publicly visible at the dashboard: http://invid.weblyzard.com.

This dashboard is an extension of the Web intelligence dashboard of the webLyzard platform [1] with news topics, story listings and also a visualisation of detected stories over time (either episodic - development of stories over the time period - or bursts - emergence of stories within the time period). Episodes link stories over time at the granularity of hourly periods, for example in Fig. 5 it can be observed how stories developed over the 48h period of 00:00 on the 18th June 2017 until 23:59 on the 19th June 2017. Only "top" stories are labelled in the visualisation. The London mosque terrorist attack clearly dominates the second 24h period, having occurred just after midnight on the 19th June. Note that the story first occurs here at 5am, which can be explained by the local time of the event meaning that most users began posting videos related to the event in the early morning. One can see how the story remained a major topic of news video throughout the rest of the day.

An alternative burst visualisation highlights not the continuity of a story but its emergence. Here in Fig. 6 the same time period is shown. For example the largest circles for the London mosque terrorist attack are at 9-10am despite the story first being detected at 5am, easily explained by the fact that by this time the UK and Western Europe was awake and reacting to hearing the story by posting content related to it. The story continues to be the main discussion point throughout the day, until the early hours of the 20th July when the news breaks of the death of Otto Warmbier.

With a story selected, the user can explore only those documents (videos posted on social networks) related to that story. Fig. 7 shows the top videos returned for the story ATTACK + LONDON MOSQUE + VAN. Videos can be played back within the dashboard and we plan eventually to enable the user to request further verification of the video material by InVID verification services through a button on the interface (to launch InVID's media verification app).

6 CONCLUSIONS

This paper has presented a novel approach to detecting breaking news and retrieving relevant user-generated video content from social networks. An evaluation of both approaches - story detection and news video retrieval - documents very promising results. The SNOW2014 data challenge has shown that comparative systems' precision and recall is not high for the domain of first story detection, as it is challenging to determine a ground truth for daily or hourly news. While other work also has demonstrated breaking news detection out of Twitter streams, our approach uniquely considers additionally the appropriate labelling of each news story for the purpose of information retrieval. The result is a data collection pipeline for the InVID platform that currently collects about 3 000 news videos on a daily basis, integrating relevant content from YouTube, Vimeo, DailyMotion and Twitter Video. Based on automatically generated story labels, we demonstrate that high content relevance can be achieved while maximising the breadth of different news stories. This data becomes immediately useful to end users via the InVID dashboard, where journalists or newsrooms can explore user-generated content around breaking news stories.

[1] http://www.weblyzard.com

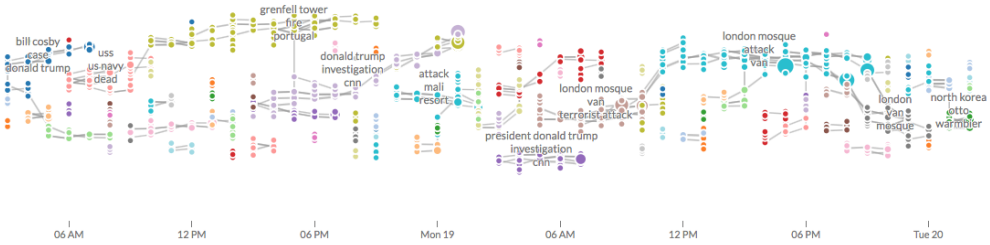

Figure 5: Episodic visualisation of stories 18-19 June 2017

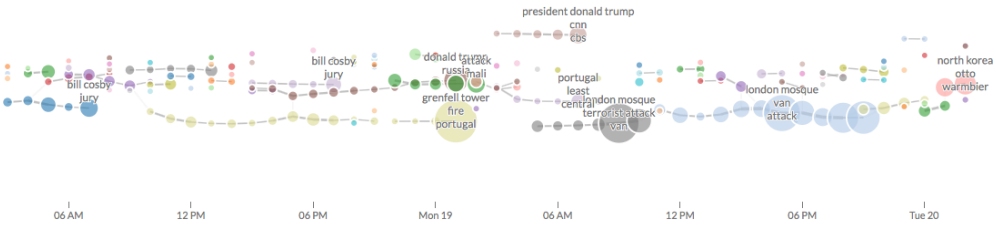

Figure 6: Burst visualisation of stories 18-19 June 2017

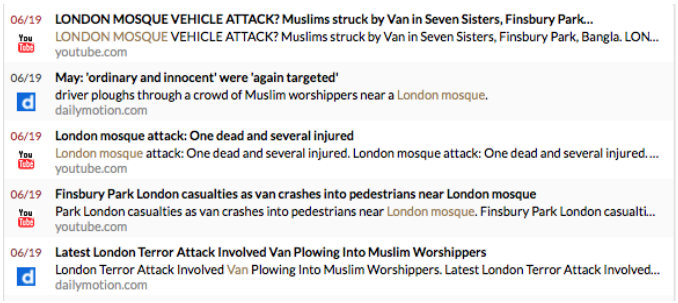

Figure 7: Videos collected for one story on 19 June 2017

To further increase the usefulness of this work in the context of supporting the multimedia verification process, we plan to explore approaches to measure the "authoritativeness" of video documents using feature-based methods such as user profile characteristics or video metadata characteristics. Currently, the video collection does not discriminate by user account, whereas a video from BBC World may be more trustworthy that one posted by a politically motivated propaganda account. This would allow users to order videos by how authoritative they may be, where less authoritative but novel video around a news story is the priority for verification. We will also explore how to model news stories in a more structured semantic model, which can also further improve news video retrieval as well as support richer ways to browse and filter news video in the dashboard, e.g. by detecting the location of a news story, we can look for videos being posted at the time of the event from around that location. Annotating videos by location (based on the news story tha they are associated with) and providing interactive means to browse the video collection along the geographic dimensions ensures high relevance especially in the case of local and regional events.

ACKNOWLEDGMENTS

The work presented in this paper is supported by the InVID research project, which has received funding from the European Unionfis Horizon 2020 Research and Innovation Programme (H2020-ICT-2015) under grant agreement No. 687786.

REFERENCES

[1] L.M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. 2013. Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia* 15, 6 (Oct. 2013), 1268–1282. DOI: http://dx.doi.org/10.1109/TMM.2013.2265080

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[3] Igor Brigadir, Derek Greene, and Pádraig Cunningham. 2014. Adaptive Representations for Tracking Breaking News on Twitter. *CoRR* abs/1403.2923 (2014). http://arxiv.org/abs/1403.2923

[4] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining (MDMKDD '10)*. ACM, New York, NY, USA, 4:1–4:10. DOI: http://dx.doi.org/10.1145/1814245.1814249

[5] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JAsIs* 41, 6 (1990), 391–407.

[6] Efthimis N. Efthimiadis. 1996. Query Expansion. *Annual Review of Information Systems and Technology* 31 (1996), 121–187. http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html

[7] Ahmed Elbagoury, Rania Ibrahim, Ahmed Farahat, Mohamed Kamel, and Fakhri Karray. 2015. Exemplar-Based Topic Detection in Twitter Streams. In *Ninth International AAAI Conference on Web and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10533

[8] Toshiaki Fujiki, Tomoyuki Nanno, Yasuhiro Suzuki, and Manabu Okumura. 2004. Identification of bursts in a document stream. In *First International Workshop on Knowledge Discovery in Data Streams (in conjunction with ECML/PKDD 2004)*. Citeseer, 55–64.

[9] Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. 2012. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2751–2754.

[10] Georgiana Ifrim, Bichen Shi, and Igor Brigadir. 2014. Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering.. In *SNOW-DC@ WWW*. 33–40.

[11] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25, 2-3 (Jan. 1998), 259–284. DOI: http://dx.doi.org/10.1080/01638539809545028

[12] Carlos Martin, David Corney, and Ayse Goker. 2015. Mining Newsworthy Topics from Social Media. In *Advances in Social Media Analysis*, Mohamed Medhat Gaber, Mihaela Cocea, Nirmalie Wiratunga, and Ayse Goker (Eds.). Number 602 in Studies in Computational Intelligence. Springer International Publishing, 21–43. http://link.springer.com/chapter/10.1007/978-3-319-18458-6_2 DOI: 10.1007/978-3-319-18458-6_2.

[13] Carlos Martin and Ayse Göker. Real-time topic detection with bursty n-grams: RGUfis submission to the 2014 SNOW Challenge. (????).

[14] Sean Moran, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2016. Enhancing First Story Detection Using Word Embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 821–824. DOI: http://dx.doi.org/10.1145/2911451.2914719

[15] Mor Naaman. 2012. Social Multimedia: Highlighting Opportunities for Search and Mining of Multimedia Data in Social Media Applications. *Multimedia Tools Appl.* 56, 1 (Jan. 2012), 9–34. DOI: http://dx.doi.org/10.1007/s11042-010-0538-7

[16] Symeon Papadopoulos, David Corney, and Luca Maria Aiello. 2014. SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media.. In *SNOW-DC@ WWW*. 1–8.

[17] Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2014. Two-level Message Clustering for Topic Detection in Twitter.. In *SNOW-DC@ WWW*. 49–56.

[18] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2011. Linking Online News and Social Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 565–574. DOI: http://dx.doi.org/10.1145/1935826.1935906

[19] Svitlana Vakulenko, Lyndon J. B. Nixon, and Mihai Lupu. 2017. Character-based Neural Embeddings for Tweet Clustering. *CoRR* abs/1703.05123 (2017). http://arxiv.org/abs/1703.05123

[20] Steven Van Canneyt, Matthias Feys, Steven Schockaert, Thomas Demeester, Chris Develder, and Bart Dhoedt. 2014. Detecting newsworthy topics in Twitter. In *Data Challenge, Proceedings*. Seoul, Korea, 1–8.

[21] Henning Moberg Wold and Linn Christina Vikre. 2015. Online News Detection on Twitter. (2015).