Geoparsing and Geosemantics for Social Media: Spatio-Temporal Grounding of Content Propagating Rumours to support Trust and Veracity Analysis during Breaking News

STUART E. MIDDLETON, University of Southampton IT Innovation Centre*
VADIMS KRIVCOVS, University of Southampton IT Innovation Centre

In recent years there has been a growing trend to use publically available social media sources within the field of journalism. Breaking news has tight reporting deadlines, measured in minutes not days, but content must still be checked and rumours verified. As such journalists are looking at automated content analysis to pre-filter large volumes of social media content prior to manual verification. This paper describes a real-time social media analytics framework for journalists. We extend our previously published geoparsing approach to improve its scalability and efficiency. We develop and evaluate a novel approach to geosemantic feature extraction, classifying evidence in terms of situatedness, timeliness, confirmation and validity. Our approach works for new unseen news topics. We report results from 4 experiments using 5 Twitter datasets crawled during different English-language news events. One of our datasets is the standard TREC 2012 microblog corpus. Our classification results are promising, with F1 scores varying by class from 0.64 to 0.92 for unseen event types. We lastly report results from two case studies during real-world news stories, showcasing different ways our system can assist journalists filter and cross check content as they examine the trust and veracity of content and sources.

Categories and Subject Descriptors: **H.3.1 [Content Analysis and Indexing]**: Linguistic processing; **I.2.7 [Artificial Intelligence]:** Natural Language Processing - Text analysis

General Terms: Algorithms, Measurement, Performance, Design, Experimentation

Additional Key Words and Phrases: Geosemantics, Geoparsing, Trust, Credibility, Veracity, Social Media, News, Breaking News, Rumours, Journalism

ACM Reference Format:

Stuart E. Middleton, Vadims Krivcovs, 2015. "Geoparsing and Geosemantics for Social Media: Spatio-Temporal Grounding of Content Propagating Rumours to support Trust and Veracity Analysis during Breaking News" *ACM Trans. Information Systems.* DOI:xxxx

1. INTRODUCTION

In recent years there has been a growing trend for the use of publically available social media content (e.g. Twitter, YouTube, Facebook, Instagram) for analytics within the field of journalism. With social media content freely available and updated in real-time breaking news journalists are turning to it to discover trending topics, on the spot incident reports and eyewitness image / video content. Often images and videos are uploaded by people on the scene before a local journalist arrives at an event to physically verify the story. Breaking news has tight reporting deadlines, measured in minutes not days, with the need to be the first to publish a breaking news story directly competing with the need to verify content [Silverman 2013;

This work is part of the research and development in the REVEAL project (grant agreement 610928) and SENS4US project (grant agreement 611242), supported by the 7th Framework Program of the European Commission
Contact author's address: Stuart E. Middleton, IT Innovation Centre, Gamma House, Enterprise Road, Southampton, SO16 7NS, UK, sem@it-innovation.soton.ac.uk

DOI:http://dx.doi.org/10.1145/0000000.0000000

Spangenberg and Heise 2014] for its credibility and truthfulness. As such journalists are looking at automated content analysis to pre-filter large volumes of social media content prior to more traditional manual verification techniques such as cross-checking and direct attempts to contact sources.

Current tools available to journalists in busy newsrooms are broadly categorized as dashboard and in-depth analytic tools. Dashboards display filtered traffic volumes, rank popular content URL's, filter content by topic or author and map geotagged content for subsequent manual retrieval. In depth analysis tools support technologies such as sentiment analysis, social network graph visualization and topic tracking. These tools are helping journalists manage social media content but there remains a big challenge when trying to identify credible and trustworthy social media content from large volumes of incoming real-time traffic. Unverified rumours and fake news stories are becoming both increasingly common [Silverman 2015] and increasingly difficult to spot. The uptake of social media is increasing all the time and both organizations and governments, potentially with vested interested in propagating false rumours, are become increasingly tech-savvy. This up-scaling of available content is in stark contrast to current best practice for journalistic user generated content (UGC) verification [Silverman 2013], which follows a hard to scale manual process involving journalists reviewing content from trusted sources with the ultimate goal of phoning up authors to verify specific images / videos and then asking permission to use that content for publication.

In the REVEAL project we are developing a real-time situation assessment framework for journalists during breaking news events. The aim is to allow journalists to filter and visualize large volumes of relevant social media content, in real-time and under breaking news timescales, to ultimately help them in the content verification process. This framework allows journalists to use social media content as a pool of crowd sourced news reports, exploiting both the 'wisdom of the crowd' to highlight popular breaking stories and identifying 'black swan' outlier content that might help reveal a deeper truth about a news story. We are working with a German national news provider to provide real-world case studies and opportunities to conduct user trials of prototype versions of our system.

Our system uses a scalable approach to geoparse text, spatially and temporally grounding real-time social media content relevant for breaking news stories. This geoparsing approach extends an innovative named entity matching algorithm described in previously published work [Middleton et al. 2014]. We describe our work extending the scalability of this approach, geoparsing location sets in parallel on an Apache Storm cluster and using a local planet deployment of OpenStreetMap to remove the need for remote geocoding. Our scalable approach can geoparse on-demand many focus areas with 100,000's of location entities under breaking news timescales and visualize real-time social media content clustered by location. This ultimately helps journalists find spatially grounded text / images / videos to help analyse facts & rumours associated with their news story.

In addition to spatially grounding content we want to be able to extract geosemantic feature properties from social media incident reports to allow a deeper analysis and intelligent filtering of the content presented to journalists. The term 'geosemantics' [Lieberman and Goad 2008] is the study of context in relation to spatial data; concretely for our work we mean contextual text relating to geoparsed locations in social media content. Our goal is to enable spatial and temporal filters so the volume of content requiring manual verification is reduced without losing any of the key information. We describe our novel approach to extracting geosemantic

features from geoparsed textual content. We focus on features relating to how the location is being referred to, such as if reports are being made insitu, in the past or represent a denial of an event at a location. We chose our geosemantic classes 'situatedness', 'confirmation', 'timeliness' and 'validity' after discussions with journalists about the verification process. These classes are useful when classifying content such as eyewitness reports, debunking reports for rumours and live news reports. Our approach is inspired by relational extraction and geosemantic analysis, using a supervised learning algorithm to classify new unseen content. We show empirically that our approach can work on any type of breaking news event (e.g. war, politics, entertainment, natural disasters) regardless of if it has been seen before or not.

We provide a set of real-time visualizations that aim to help journalists navigate through large volumes of social media content items. A ranked item view provides journalists with a list of social media content items ranked according to a set of dynamically adjustable filter criteria. A temporal content view shows a similar list ranked by timestamp. Finally a geospatial view allows content clustered by location to be displayed on a map, with content items / URI's / tags for each location available to be seen by journalists when clicked upon. Our aim is to support spatio-temporal content grounding for rumour analysis. The temporal view allows journalists to trace rumours back in time to discover their source. The spatial view allows journalists to cross-check content spatially and corroborate facts and locations from eyewitness reports at the time of a rumoured event.

Finally we report two case studies based on real-world news events. The first case study looks at the false rumour that the New York Stock Exchange (NYSE) was flooded during the Hurricane Sandy in October 2012. This rumour was started by the US Weather Channel and propagated by the US news provider CNN. We look at how effective spatial and temporal filtering of content is during the height of the rumours, when TV and social media was awash with conformations and denials. We show that temporal filtering of content spatially grounded to the NYSE's location can significantly reduce content volumes without loss of the key content journalists used at the time to uncover the truth behind this story. The second case study looks at the fall of Donetsk airport in Ukraine in early January 2015, and the conflicting claims made by Ukraine and Russia TV regarding who controlled the airport on the 20th January 2015. We look at how spatial clustering of content filtered using a temporal window of interest could be used to help journalists identify key images and videos from social media. We find that several of the You Tube videos highlighted by our Donetsk airport location cluster also appear in reports published by Russian news provider Life News at that time.

In section 3 we describe our framework for scalable real-time situation assessment. Sections 4 and 5 then describe briefly our approach to geoparsing and geosemantic feature extraction. Section 6 describes our spatio-temporal visualization work with an example of a geospatial view for the Ukrainian case study. Section 7 has results from an empirical evaluation of our geosemantic feature extraction and section 8 discussed the findings in more detail. We finish with results from our two case studies in section 9 and conclusions in section 10.

This paper makes a number of contributions to the state of the art. The scalability enhancements to our geoparsing algorithm represent an advancement of our previously published state of the art named entity matching approach. The geosemantic feature extraction method is a novel contribution, exploiting an innovative wildcard phrase formulation inspired by both relational extraction and

geosemantic analysis. Finally the evaluation of these two techniques and application to real-world case studies provides a contribution for practitioners wishing to understand the effectiveness of spatio-temporal content grounding for the purposes of rumour detection.

2. RELATED WORK

2.1 Commercial tools

There are a number of commercial tools available today [Spangenberg and Heise 2014] to support journalists in managing social media content. Dashboard applications (e.g. Tweetdeck, Sulia, Storyful, Flumes, WebLyzard) allow journalists to track news stories, alerting them to new and relevant content, trending topics and influential people. These dashboard applications allow journalists to drill down into content and get contact details for a subsequent manual verification process (e.g. verification via a phone call to the content author). For in depth analysis there are tools supporting sentiment analysis (e.g. Bing Elections, SocialMention), social network graph visualization (e.g. MentionMapp, Bottlenose) and topic tracking (e.g. Trackur). Some research systems [Raz et al. 2015] utilize crowdsourcing to classify content as news worthy. There are also tools such as Geofeedia which will display geotagged social media content interactively on a map. Recent news visualization systems [Samet et al. 2014] are moving beyond keyword searches for social media content and offering new representations such as spatial browsers. Our spatio-temporal visualizations are motivated by work in this area.

2.2 Journalistic verification

A good description of journalistic practice for verification of user generated content can be found in [Silverman 2013]. This handbook outlines a set of case studies with examples from organizations such as BBC News, GuardianWitness and Storyful. The approach journalists follow is a manual one, based on source identification (e.g. phoning up content authors), content identification (e.g. finding out the location, time and date of content), cross-referencing between different reports (e.g. eyewitness reports from different sources) and looking to obtain permission to use content from the author / originator. For more in-depth analysis investigation teams such as Bellingcat have provided how-to guides [Higgins 2014] for manual verification activities such as geolocating videos. New methodologies are also emerging to address the viral nature of rumours in social media [Silverman 2015], however even these refined processes are still manual in nature and tooling limited to existing dashboards and in-depth analytic tools. Our work is focussed on adding some automation to make the manual part of this process more efficient.

2.3 Location extraction

The extraction of location from text is called geoparsing, and this has been well studied in the field of natural language processing [Gelernter and Mushegian 2011]. Approaches to geoparsing are usually based on either named entity recognition (NER), using annotations such as parts of speech (POS) tags, or named entity matching (NEM) using a gazetteer of known locations. Named entity recognition of locations for micro-blog information is challenging due to the short text length and wide variety of grammatical styles. Typical approaches include conditional random fields (CRF) coupled with named entity recognition [Ritter et al. 2011] and entity disambiguation using a reference corpus such as DBpedia [van Erp et al. 2013].

Other approaches include entity disambiguation with machine learning techniques such as Expectation-Maximization [Davis et al. 2012]. Our named entity matching based approach to geoparsing has been shown [Middleton et al. 2014] to have a better precision than state of the art named entity recognition techniques without compromising F1 scores.

2.4 Geosemantics

Geosemantics [Lieberman and Goad 2008] is a relatively recent term and covers any contextual information in text relating to a location, including position refinement of location references and any time or sentiment associated with incidents at the reported location. After geoparsing is performed relational extraction techniques [Bontcheva et al. 2013] can be used to achieve position refinement of the geoparsed location. Existing work on position refinement has mostly focussed on spatial role learning [Kordjamshidi et al. 2012; Bastianelli et al. 2013], where phrases for 'trajectors' (i.e. position modifier) and 'landmarks' (i.e. location) are extracted and classified. This is helpful when references to a location are augmented with phrases such as '5 miles north of London'. Our 'situatedness' geosemantic classification is somewhat different to this type of position refinement as we are interested in eyewitness reports not position refinement of the reported location. Our approach could be used alongside position refinement techniques.

No prior work has been published using our choice of geosemantic classes. As such a direct benchmark against previous work is not possible. We do however evaluate our work using the standard TREC-2012 corpus [Soboroff et al. 2012] and this allows future researchers to easily benchmark directly against the results in this paper.

2.5 Temporal extraction

The extraction of time references from text is called temporal expression extraction [Verhagen et al. 2010]. These techniques seek to extract the time references from multi-lingual patterns within text (e.g. '30 minutes ago'). Popular techniques for temporal expression extraction include named entity recognition coupled with rule-based temporal heuristics [Grover et al. 2010], conditional random fields [Llorens et al. 2010] coupled with a model for temporal semantics, and linguistic rule-based approaches [Teresa et al. 2010]. We try to extract a concept of time from text by using our 'timeliness' class, but we are only interested in a broad past / present / future categorization for filtering purposes. Our supervised learning approach to classify timeliness could be used alongside work on temporal expression extraction.

2.6 Sentiment analysis

The field of sentiment analysis for text is a mature one [Feldman 2013], aiming to extract concepts like good/bad and positive/negative and data mine opinion from application domains such as political discourse and the launch of new consumer products. A wide range of techniques have been used working at the document level down to sentence and aspect (i.e. attribute) levels. At document level supervised approaches, applied to text representations such as bag-of-words [Pang et al. 2002], and unsupervised approaches, based on a lexicon of phrases [Taboada et al 2011] and parts-of-speech patterns [Turney 2002], have all worked well. Examples of sentence level work include supervised learning binary classifiers [Yu and Hatzivassiloglou 2003] and at aspect level techniques such as noun phrase frequency thresholding [Minqing Hu and Bing Liu 2004]. Understanding sarcasm [Tsur et al. 2010] has

proved a particular challenge for sentiment analysis which we also found was a factor in the false positive rate we see relating to our geosemantic 'confirmation' class. In some respects our 'confirmation' class could be thought of as a geospatially specialized type of sentiment analysis.

2.7 Rumour detection and visualization

Recent work on rumour detection from microblog text has focussed on using both linguistic patterns and features from mentioned images. For example tweeted claims of fake images and debunking reports [Zhao et al. 2015] can be identified and clustered together for relevance to verification. Relevance results up to 0.7 precision for top 10 content lists [Zhao et al. 2015] can be achieved this way for news events like the Boston bombing 2013. Source attribution patterns [Middleton 2015] can be used to rank claims in order of trustworthiness, allowing a high precision between 0.94 and 1.0 at the expense of a lower recall between 0.43 and 0.72. Forensic image analysis can also be used to provide evidence for faking, with [Boididou et al. 2015] using a combination of JPEG compression features (e.g. discrete cosine transform coefficients) and Exchangeable image file format (Exif) metadata to train supervised classifiers in addition to linguistic features. This approach is more robust overall with a precision of 0.86 and recall of 0.96.

Systems such as [Carton et al. 2015] [Finn et al. 2014] are visualizing content and authors involved in propagating known rumours over time via timelines and social network graphs. These visualizations allow semi-automated interactive analysis with questions such as who are the individuals who have the biggest impact on a specific rumours explored. Other work [Zhao 2014] has examined overlaying social network interconnections to temporal graphs of rumour retweets, revealing active users in both graphs during propagation periods as a rumour goes viral. Our work on geosemantic features and spatio-temporal visualization through map-based visualizations is quite complimentary to these other approaches, and could easily work in conjunction to them.

3. ARCHITECTURE

The REVEAL project's scalable real-time situation assessment framework is built upon the Apache Storm [1] distributed real-time computation system and the RabbitMQ[2] Advanced Message Queuing Protocol (AMQP) message bus. It is scalable to large throughputs of real-time social media content supporting analysis of many breaking news stories at once. The overall framework architecture is shown in Fig. 1.

A real-time crawler streams social media content from the Twitter Streaming API, Twitter Search API, YouTube Search API and Instagram Search API to a RabbitMQ message bus for geoparsing and geoclassification. The use of RabbitMQ provides a scalable message backbone that can handle the throughput of messages expected across the entire processing pipeline. Journalists can make additional search and stream requests as a story unfolds, allowing keywords and hashtags that start trending to be added.

We deploy our geoparsing and geoclassification services as Storm topologies so they can be instantiated on a computing cluster. New focus areas can be geoparsed on demand. Each block of new locations are first pre-processed from a local Planet

[1] http://storm.apache.org/
[2] http://www.rabbitmq.com/

OpenStreetMap [3] database and then loaded into memory by a Storm geoparse topology instance. We can scale up the number of locations processed simply by adding more machines to the Storm cluster. The geoclassification is also run as a Storm topology and scales in the same way. All geoparse and geoclassification results are published to RabbitMQ for aggregation into a common situation assessment suitable for journalists.

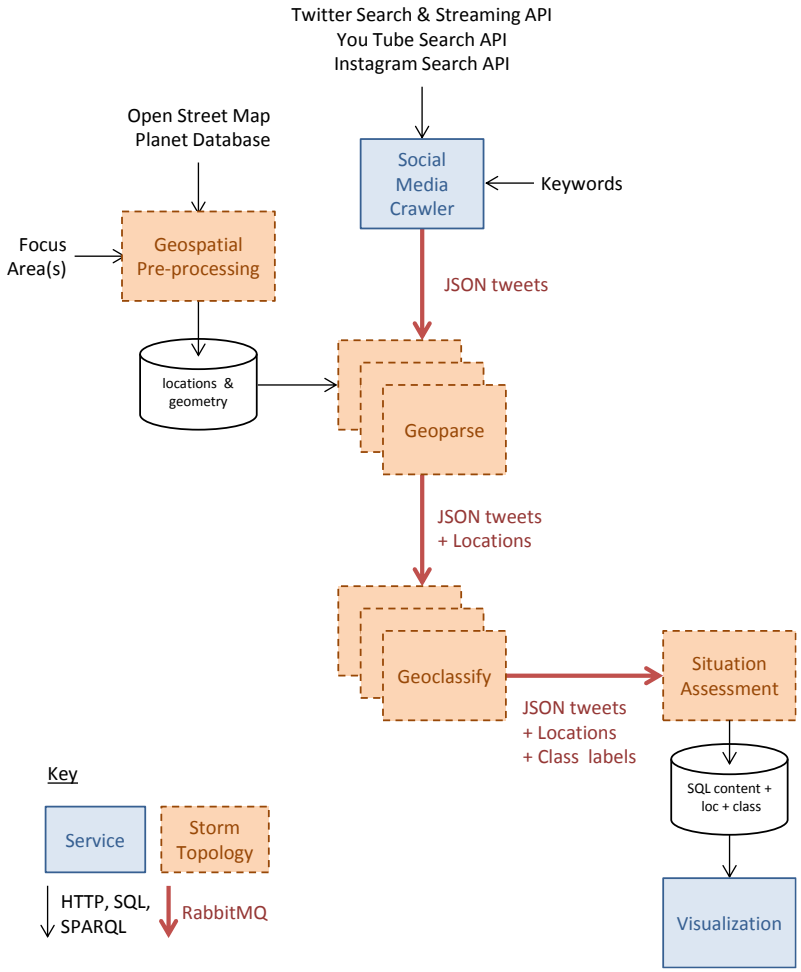

Fig. 1. Scalable architecture for geoparsing and geosemantic analysis of social media content

A situation assessment service aggregates content annotations (i.e. geoparsed location and geosemantic class labels) and maintains a real-time database with live situation assessments. This can be visualized on demand, allowing journalists to browse filtered and clustered sets of social media content interactively, rending them on geospatial maps and ranked item views.

4. GEOPARSING AT SCALE

Our geoparsing Storm process represents an extension of a state of the art named entity matching algorithm (NEM) first published by [Middleton et al. 2014]. We

[3] http://www.openstreetmap.org/

briefly summarize our approach in this section then describe in more detail how we have extended its scalability through parallelization on an Apache Storm cluster and removed the need for remote geocoding via a local planet deployment of OpenStreetMap.

The geoparsing approach we use is based on named entity matching and requires access to a pre-loaded map database. We use a set of pre-loaded global cities (i.e. 300,000 locations) plus a number of focus areas with region, street and building information (e.g. New York). Pre-processing includes a token expansion step where abbreviations and language specific variants of street and building types are used to create token sets that best represent the way social media users refer to each location. Real-time text is tokenized into n-gram tokens and matched to an in-memory cache of known location n-gram tokens. The use of pre-loaded location data also allows us to avoid error prone named entity recognition (NER) steps and has been shown [Middleton et al. 2014] to yield higher precision results (i.e. 0.9 or higher) compared to state of the art NER approaches without compromising on the overall F1 scores (i.e. 0.8 or higher).

Most geoparsing services either use a remote geocoding service such as the Google Geocoder[4] or location name lookup in a gazetteer such as Geonames[5]. Remote geocoding does not scale well as all remote geocoding services have rate limits (e.g. 10 requests per second, 100,000 requests per day for Google Geocoder). This is insufficient for real-time work where multi-stream throughputs of up to 50 content items per second are typical. Gazetteer lookup has no such rate limits but typically only works at a region level, with access to street and building details not available.

We use a local planet deployment of OpenStreetMap's database, based on PostgresSQL & PostGIS technology, to avoid geocoding and allow us to service on-demand focus areas requests from anywhere on the globe. When a new focus area is requested for a breaking news story a pre-processing service queries the OpenStreetMap database to retrieve all administrative regions, streets and buildings in that area. The location names are then tokenized and these tokens are expanded to include abbreviations and name variants suitable for matching to microblog text (e.g. london street, london st.). Super-region identifiers for all locations are computed using a PostGIS geospatial query (e.g. the region Donetsk is geospatially contained by its super-region Ukraine). This pre-processing task takes minutes to complete depending on the number of locations in each focus area.

Once pre-processed, blocks of locations are cached into memory by real-time geoparse processes. If a focus area has a lot of locations it is chunked into blocks of 100,000 and the geoparsing parallelized. Each geoparse process receives a real-time stream of JSON-formatted social media content, tokenizes it and matches these tokens to the in-memory cached location entities. An aggregation process takes location annotations for individual content items and applies a set of location name disambiguation heuristics. Name disambiguation is important since location names are often found repeated across the globe (e.g. Donetsk, Ukraine and Donetsk, Russia). If available we use content geotags to add confidence to possible location matches nearby. We also add confidence to possible location matches were super regions or nearby locations are mentioned in the text after the original location mention. The set of possible location matches is then ranked by confidence and the highest confidence matches selected.

[4] http://developers.google.com/maps/documentation/geocoding/
[5] http://www.geonames.org/

The accuracy of our geoparsing algorithm is not the primary focus of this paper, and has been analysed in depth in [Middleton et al. 2014]. We have shown previously that for major news events, such as natural disasters, geoparsed social media maps can mirror well human authored expert damage assessments. The map on the left of Fig. 2 is the official Storm surge map created after the New York Hurricane Sandy flooding in 2012 by human analysis of aerial and satellite imaging. The map on the right of Fig. 2 is a 5 day Tweet flood map created during the same event based only on tweeted incident reports. Our new work on spatio-temporal visualization in this paper builds upon our earlier published work, extending and tailoring it for use by journalists working on different types of news stories.

With regards scalability we have seen [Middleton et al. 2014] typical peak throughput from the sampled Twitter Streaming API during major events (e.g. flooding keyword filtered streams during Hurricane Sandy 2012) of 5 content items per second. This would be more if Twitter firehose access was available. A typical application might receive 5 to 10 different keyword filtered streams so a geoparsing target throughput of between 25 to 50 content items per second is reasonable. Our geoparsing algorithm can process 11 content items per second (with 310,000 global cities & regions matched) with 1 process running on a single computing node, fully loading 1 CPU core. Our throughput rates increase linearly when we add compute nodes to our cluster since the geoparsing and geosemantic classification is designed to be naively parallelizable. On our project testbed we have 4 compute nodes, each with 8 CPU cores, running 4 geoparse processes each. We can therefore geoparse with a throughput of up to 176 content items per second. We find about 20% of news event keyword filtered tweets typically contain a location reference so a reasonable target for geosemantic classification throughput is between 5 and 10 content items per second (i.e. 20% of the raw content throughput). Our geoclassification algorithm, pre-trained with top 100 features, can classify 60 content items per second with 1 process running on a single compute node. Out cluster-based deployment therefore delivers good scalability that can handle the demands of our journalism use cases.

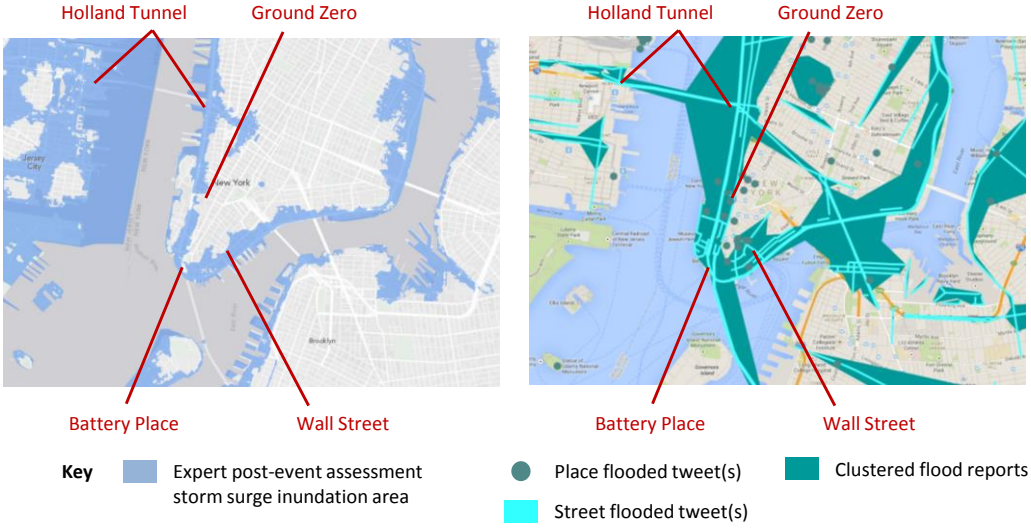

Fig. 2. Crisis map comparison for New York's 2012 flooding [Middleton et al. 2014]. Left image is the ground truth post-event US National Geospatial Agency (NGA) impact assessment showing storm surge inundation. Right image is a 5-day tweet flood map. Red annotations show major flood locations.

5. GEOSEMANTIC FEATURE EXTRACTION

Our approach to geosemantic feature extraction from microblog text is motivated by coupling concepts from relational extraction and geosemantics. In relation extraction text positioned between named entities, for example a person and a location, allows possible relationships to be identified. In geosemantics the focus is on text associated with a location to make the report more exact, for example '10 miles north of London'. We are interested in textual relations associated with a location but not necessarily connected to another named entity. For example 'eyewitness report in London' would indicate the geosemantic feature class 'situatedness'.

We first extract contextual terms (i.e. words) within a certain semantic distance (i.e. word distance left or right) of a geoparsed location term. These contextual terms are then used to create feature phrases, ranked according to their discriminating power, which is then used to train a supervised learning classifier. A pre-trained classifier can then process streams of tokenized microblog content in real-time. For example a sentence "My brother saw a lot of flooding in London yesterday." and a semantic distance of 6 tokens would yield a text fragment for analysis of "saw a lot of flooding in London yesterday.".

We use the Python Natural Language Toolkit (NLTK) [Bird et al. 2009] to clean and tokenize UTF-8 microblog text. We use the Punkt word and sentence tokenizer and apply weak stemming (i.e. plural removal) as location terms are sensitive to stronger stemming. Contextual terms are taken from the left and right of geoparsed location terms at a semantic distance. The optimal semantic distance was empirically determined (see section 7) with 12 terms a robust distance choice.

A novel aspect of our approach is our feature phrase representation that allows any combination of terms, parts of speech tags and wildcards. During analysis of our tweet datasets we observed that the way locations were referred to (i.e. grammar and tense) appeared more important than the actual words used. For example insitu reports generally used the active voice as the reporter was on the scene engaged directly with the event. It was also observed that microblog text lacked rigour in terms of sentence construction, so capitalization of words or basic sentence construction could not be relied upon and additional terms, such as hyperbole and emoticons, often appeared as 'padding' between a key contextual phrase and the location token it was referring to. For example a confirmation report might say 'flooding reported in wall street' or it might say 'floods are getting really bad now down in wall st.'.

Feature phrases are made up of n-gram phrases calculated from all linear combinations of contextual terms. An n-gram phrase is simply a tuple with a sequence of terms / tags of length N. Since our feature phrases encode sentence structure we can safely use a 'bag-of-words' representation for the supervised learning stage without losing important information about word sequencing.

Parts of speech tagging is the process of annotating terms based on lexical categories (e.g. NN tag for Noun). We used TreeTagger[6] for parts of speech tagging as it is a mature tagger that supports many languages and has good community support. For every feature phrase we calculate all combinations of phrases mixing terms, parts of speech tags and wildcard operators. We also support the Stanford POS tagger[7] but found TreeTagger more useful as it supports a wider range of European languages such as Russian.

[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[7] http://nlp.stanford.edu/software/tagger.shtml

Going one step further we provide optional named entity recognition tags to augment parts of speech tagging. Named entity recognition classifies tagged text to identify likely terms relating to people, organizations and locations. We used NLTK's inbuilt maximum entropy named entity classifier trained on the ACE English text corpus. Algorithm 1 shows the final feature phrase calculation algorithm along with example 1 of what the final feature phrases look like for a typical tweet.

ALGORITHM 1. calc_feature_phrase

Input: tagged_sentence
Output: feature_set

```
feature_set = []
min_gram = 2; max_gram = 5

if ENABLE_GEOTERM then
  replace loc term(s) with '#loc#' in tagged_sentence
if ENABLE_GEOPOS then
  replace loc tag(s) with 'LOC' in tagged_sentence
if ENABLE_NE then
  exec named entity recognition (NLTK max entropy approach)
  replace tag(s) with NE labels in tagged_sentence

for nTermPos = 1 to len( tagged_sentence ) do
  token_seq = tagged_sentence[ nTermPos : END ]
  for gram = min_gram to max_gram do
    phrase_set = tuple( token_seq[nTermPos : nTermPos + gram ] )
    phrase_set.add( all wildcard versions of phrase )
    phrase_set = calc_mixed_term_pos_phrases( phrase_set )
    feature_set.add( phrase_set )
  endif
endif
return feature_set
```

EXAMPLE 1. FEATURES FOR A TORNADO TWEET

Tweet = "Oklahoma tornado filmed by Newcastle resident"
Geoparse = term Oklahoma, start_index 1, end_index 1
POS = Oklahoma/NP tornado/NN filmed/VVN by/IN Newcastle/NP resident/JJ
Feature Set [3gram] =
(Oklahoma tornado filmed), (tornado filmed by), (filmed by Newcastle) ...
(NP tornado filmed), (Oklahoma NN filmed), (Oklahoma tornado VNN), (Oklahoma NN VNN), (NP tornado VNN) ...
(Oklahoma * filmed), (Oklahoma * by), (Oklahoma * Newcastle), (Oklahoma * resident) ...
(NP * filmed), (Oklahoma * VNN), (NP * by), (Oklahoma * IN), ...

To facilitate an understanding of which feature types work best we made the algorithm capable of providing any combination of features using plain words (TERM), parts-of-speech (POS) and named entities (NE). We provided a setting (GeoTERM) where location terms were simply replaced with the word '#loc#' to see if

geoparsed location tagging could be used, avoiding the need for parts of speech tagging. Lastly we provided a setting (GeoPOS) where location tags were replaced with a tag 'LOC', avoiding the need for named entity recognition.

This feature calculation algorithm results in a large combination explosion of possible phrases. The next step is therefore a two phased feature selection process to choose the most appropriate features to be used with our classifier. Algorithm 2 shows this process. Phase one of feature selection involves filtering all features phrases that are not common across the training event datasets for each class. This ensures event-specific terms like 'flooding' do not appear which would be inappropriate for different event types like a tornado. Phase one selection typically leaves about 20,000 features phrases. Phase two selection applies a TF-IDF algorithm to identify, from the remaining feature phrases, which ones are most highly discriminating. Equation (1) shows the TF-IDF algorithm used. We use a TF threshold of 10% of the max TF score before inclusion of a term into the DF metric since it is unlikely features never occur at all in a corpus. We select the topN features phrases per class label from a list ranked by TF-IDF score to give us a final more manageable set of features. The optimal topN feature value was empirically determined (see section 7) with a top 100 feature setting producing good results compared to the computation power needed to process these features.

Feature selection is particularly important to reduce the size of the training set required by the WEKA classifiers used next. Limiting the feature size to a top N best features (e.g. top 100) reduces the classifier memory footprint and the number of CPU cycles required to calculate each classification result. This in turn provides a scalable classification solution that balances classification accuracy against overall classification speed.

TF-IDF algorithm
Document = class corpus document of concatenated features
from training sentences
Term = feature phrase

N = number of documents
$TF_{t,d}$ = term t's frequency of occurrence in document d
DF_t = number of documents where TF_t is above 10% of max TF
$IDF_t = \log(N / DF_t)$
$TF\text{-}IDF_{t,d} = (1 + \log(TF_{t,d})) * IDF_t$ (1)

ALGORITHM 2. feature_selection

Input: feature_set, class_label_set
Output: filtered_feature_set

filtered_feature_set = []
for each class_label in class_label_set **do**
phase 1 filter : features common between event types
phase 2 filter : top N features ranked by TF-IDF score
add remaining features to filtered_feature_set
return filtered_feature_set

After feature phrase selection we generate WEKA [Witten and Frank 2011] datasets and train a WEKA supervised learning classifier. We made the classifier configurable in our experiments and we tested using Naive Bayes, J48 decision tree and IBk k-nearest neighbour classifiers. We also added a bagging, Random Forest, and boosting, LogitBoost, classifier using a base J48 decision tree. A configurable classification threshold (e.g. 0.7) is provided before a class label is accepted based on the WEKA classifiers probability score for a class label.

We chose to use 4 classes in our work: 'confirmation', 'timeliness', 'situatedness' and 'validity'. The 'confirmation' class relates to confirmation or denial of an incident at a location (e.g. 'there is no flood in New York'). The 'timeliness' class relates to if the incident at a location is being referred to in the past, future or present tense. The 'situatedness' class relates to if the reporter is actually at the scene of the event or if someone is commenting remotely (e.g. at home maybe in another country). The 'validity' class is really a noise identification class, capturing any references to a location name that is not actually a valid location at all (e.g. 'my friend Chelsea burst into floods of tears last night'). These classes were chosen after dialogue with journalists regarding what is most useful for the verification process.

6. SPATIAL-TEMPORAL VISUALIZATION FOR JOURNALISTS

In order to showcase the large volumes of geoparsed and geosemantically labelled social media content in a manageable way we provide journalists with a real-time web-based spatio-temporal visualization for each news story. We provide a ranked items view, temporal content view and geospatial map view of clustered content items. Each visualization supports multi-dimensional filters, allowing journalists to change filter criteria interactively and see those views on the data that are most appropriate to aspects of a news story the journalist is most interested in. These visualizations are currently at a prototype stage and we are working with a German national news provider to tailor them to best meet the needs of journalists.

The ranked item view is a ranked list of social media content items according to a set of dynamically adjustable filter criteria. The temporal content view is a similar list ranked by content timestamp. All geoparsed and geosemantically labelled content items are aggregated in real-time into a set of Postgres & PostGIS database tables. There is a main content item table and a set of sub-tables for mentioned locations, tags and URI's. All tables are cross indexed so, for example, a list of URI's ranked by mention frequency can be calculated for each specific location.

The interactive map view is a web-based OpenLayer[8] driven view using a backend Geoserver[9] installation. This map view allows all locations to be displayed on a map, and content items / URI's / tags displayed when each location is clicked upon. Locations can be regions (e.g. Donetsk), streets (e.g. Київський проспект) or buildings (e.g. Donetsk Airport).

Example screenshots taken from our map visualization can be seen in Fig. 3 and Fig. 7. These figures give an idea as to how the journalists can interactively explore, both spatially and temporally, grounded content for the purposes of rumour investigation. A journalist will typically start by looking at content relevant to a general area or time frame and then refine the search to 'zoom in' on more specific areas or time frames. Typically for rumour investigation this will involve looking for

[8] http://openlayers.org/
[9] http://geoserver.org/

the first mention of a rumour (i.e. its source) or reviewing sets of images / videos geospatially nearby a reported event (i.e. looking for content to corroborate a key image / video).

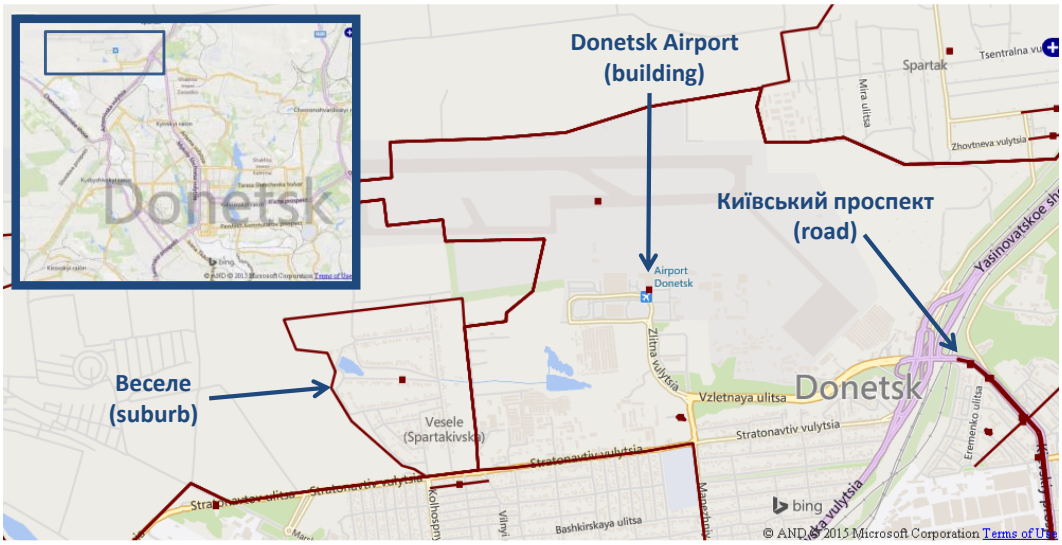

Fig. 3. Screenshot from geospatial visualization of content from Twitter, You Tube and Instagram for 20th January 2015 matching keywords for the Ukraine 2015 crisis. Red lines indicate geoparsed regions, streets and buildings. Clicking on each region, street or building brings up a list of images / videos ranked by mention frequency. Mapping courtesy of Bing Maps.

7. EVALUATION OF GEOSEMANTIC CLASSIFICATION

We conducted four experiments to examine different aspects of our geosemantic classification approach, using event datasets we crawled from Twitter and the standard TREC 2012 [Soboroff et al. 2012] microblog dataset. Table 1 outlines a detailed breakdown of classes within our manually labelled Twitter event datasets.

Manually labelling datasets is a standard approach to creating a ground truth for evaluation and is used by many researchers [Gelernter and Mushegian 2011] [Middleton et al. 2014] [Bontcheva et al. 2013]. Typically manually labelled dataset sizes range from 100's to a few 1000's of tweets and are limited by the time it takes to label the data. Providing a direct benchmark dataset for this geosemantic classification is not possible because our geosemantic classes are unique to this work. However by using the TREC dataset for one of our analysed events we allow future researchers the chance to test their systems on the same data and benchmark against this work.

The first 4 events covered were the Hurricane Sandy flooding in New York October 2012, Oklahoma tornado in May 2013, Ukraine conflict in Aug 2014 and the Scottish independence referendum in Sept 2014. These tweet datasets have been crawled by us, using the Twitter Streaming API along with event specific keywords (e.g. 'flooding'), and lasting between 1-5 days depending on the event type. The crawling resulted in 100,000's of tweets so we randomly sampled them to create a much smaller and more manageable dataset, removed near duplicate tweets to avoid any bias due to repetition, and manually labelled the remaining tweets according to our 4 class types.

Our 5th event is the Chicago Blizzard Jan 2011 which appears in the TREC 2012 dataset for event type MB57 [Soboroff et al. 2012]. These tweets were also manually

labelled by us as the TREC conference only worked at an event type level of granularity and did not use our 4 class types. Note that the TREC dataset class label coverage is not ideal, with only a few training examples for 'neg' and 'future' and none for 'insitu' labels; the effect of this is discussed later.

Before starting our 4 experiments we conducted some investigations into the optimal settings for the TF-IDF topN feature selection threshold and contextual semantic distance around location tokens. We tested topN threshold values between 25 and 500 features and found that adding more features improved the quality of the classifiers at the expense of longer computation time. We concluded that going above 100 features per class did not yield a significant advantage in terms of classification accuracy. We tried contextual semantic distances between 6 and 20 terms and found that a 12 term distance gave the best balance between capturing important contextual text whilst still removing unrelated text far away from the location reference. All 4 experiments reported in this paper used a topN threshold of 100 features and semantic distance of 12 tokens.

We also looked at different feature phrase gram sizes, manually inspecting them in terms of their suitability to the class corpus they were generated from. It is important not to use too big a range otherwise the number of combinations of feature phrases becomes too large to process quickly. We found that a feature phrase gram size range of 2 to 5 produced good results and this setting is used in all 4 experiments.

Table 1. Breakdown of event datasets used in experiments

Event Dataset	# Tweets	Class confirmation	Class timeliness	Class situatedness	Class validity
#1 Flood New York 2012	1045	724 pos 48 neg 273 na	115 past 705 present 27 future 198 na	319 remote 68 insitu 658 na	925 valid 120 na
#2 Tornado Oklahoma 2013	1045	403 pos 3 neg 639 na	65 past 780 present 12 future 188 na	279 remote 16 insitu 750 na	922 valid 123 na
#3 Conflict Ukraine 2014	1211	721 pos 8 neg 482 na	85 past 476 present 24 future 626 na	467 remote 2 insitu 742 na	983 valid 228 na
#4 Referendum Scotland 2014	1482	1081 pos 23 neg 378 na	156 past 858 present 32 future 436 na	108 remote 11 insitu 1362 na	1302 valid 180 na
#5 TREC 2012 Chicago Blizzard 2011 TREC cluster MB57	502	74 pos 2 neg 426 na	55 past 311 present 4 future 132 na	268 remote 0 insitu 234 na	289 valid 213 na

7.1 Experiment 1: Feature Phrase Comparison

The first experiment examined which feature phrase types produced the best results. We performed a 10-fold cross validation on the first 4 event datasets and looked at results for different feature phrase types across the available classes and classifiers.

We found that feature phrase types that included POS tags performed best and we conclude that POS tags alone are the most robust choice for feature phrase type. We think POS works better than TERM or GeoTERM because what matters is the grammar around contextual sentences as opposed to the exact words used when

determining how a location is being referred to. As a result we use the POS feature phrase type in all of the subsequent 3 experiments.

7.2 Experiment 2: Known Event Classification

The second experiment examined which classifier produced the best results using feature phrase POS. We performed the same 10-fold cross validation as the first experiment, using the same 4 event datasets, but this time compared results from five different WEKA classifiers using the POS feature phrase type. The chosen classifiers included a probabilistic (i.e. Naive Bayes), normal, bagged and boosted decision tree (i.e. J48, Random Forest and Logit Boost) and k-nearest neighbour (i.e. IBk) supervised learning classifiers.

To reduce false positives we only allowed results with a classifier probability of 0.7 or higher to be a true positive (TP), with less certain results recorded as a false negative (FN). This 0.7 threshold was empirically found to work well and is used in all experiments. An incorrect classification was recorded as a false positive (FP). After recording TP, FP and FN for all 10 folds we calculated the mean precision, recall and F1 scores along with standard deviations so statistical significance can be seen clearly in the results. Equation (2) shows the definition of the metrics used.

These results represent optimal performance for known event types. Results can be seen in Fig. 4. The unboosted J48 decision tree performed best overall across the 4 classes, with an F1 score of between 0.70 to 0.88, with IBk best for class 'validity' with an F1 score of 0.92. If only precision is important then Logit Boost, Random Forest and IBk k-nearest neighbour classifier performed best with precision scores between 0.76 and 0.86. This high precision comes at the expense of a low recall however, caused by many false negatives.

$$\begin{aligned} &\text{Precision } (P) = tp/(tp + fp) \quad tp = \text{true positive, } fp = \text{false positive} \\ &\text{Recall } (R) = tp/(tp + fn) \quad tn = \text{true negative, } fn = \text{false negative} \\ &\text{F1 measure} = 2*PR/(P+R) \end{aligned} \tag{2}$$

The IBk classifier was less stable across folds than other classifiers, with a maximum F1 standard deviation of 0.06. The J48 classifier was very stable with a standard deviation of 0.01. Taken across all 4 classes the superiority of the J48 classifier is statistically significant.

It should be noted that the number of training examples for individual class labels are well balanced in our datasets with the exception of the timeliness 'future' and situatedness 'insitu' labels. These have a low number of training examples in some of our event datasets, mostly because they represent relatively rare incident report types compared to the majority of social media traffic. The result for the 'future' label in isolation has a F1 score of 0.38 compared to the mean timeliness F1 score of 0.69. The 'insitu' label has a F1 score of 0.23 compared to the mean situatedness F1 score of 0.82. We highlight these outlier class labels as the cross-fold mean F1 scores hide individual class label performance. We think that training set class label imbalances can be corrected in the future with the addition of more training examples, and thus it does not affect our overall comparison of feature types or classifiers.

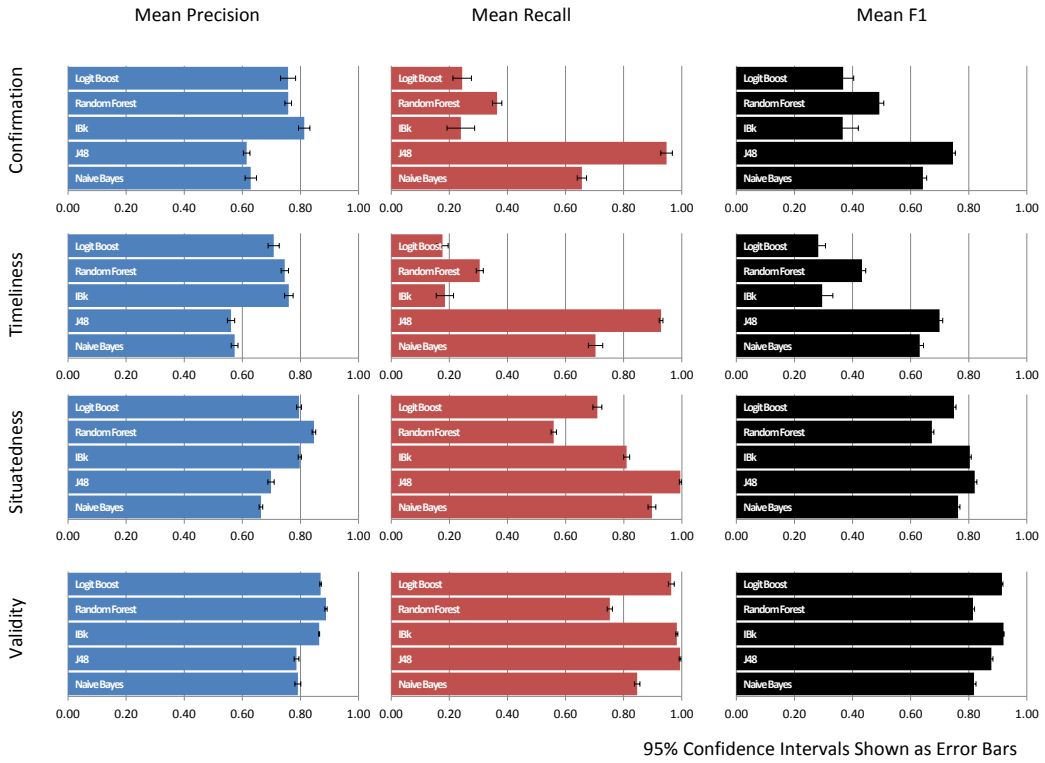

Fig. 4. 10-fold cross validation results comparing classifiers across all 4 classes using the POS feature type. These results represent performance for *known* event types. The J48 decision tree is best overall except for the class validity where IBk k-nearest neighbour is slightly better.

7.3 Experiment 3: Unknown Event Classification

The third experiment examined how the same classifier set performed when tested on an unknown event type dataset that has not been trained for. We performed a leave one out cross validation, testing using each of the first 4 event datasets in turn and training with the remaining 3 event datasets. The mean precision, recall and F1 scores were calculated alongside standard deviation. This experiment represents realistic performance, as opposed to optimal performance, since we cannot expect to have a trained event dataset for every breaking news story type that might occur in the future. The experimental conditions were identical to experiment 2 allowing results to be directly compared. The results are in Fig. 5.

The unboosted J48 decision tree again performed best overall across all 4 classes with an F1 score of between 0.64 to 0.87. With regards precision we found the Logit Boost, Random Forest and IBk were very similar again with precisions of between 0.62 and 0.89. As expected results were worse compared to experiment 2, although class valid was very similar. This is probably due to the fact that invalid location references are not event specific and so this training set is very robust to new unknown event types.

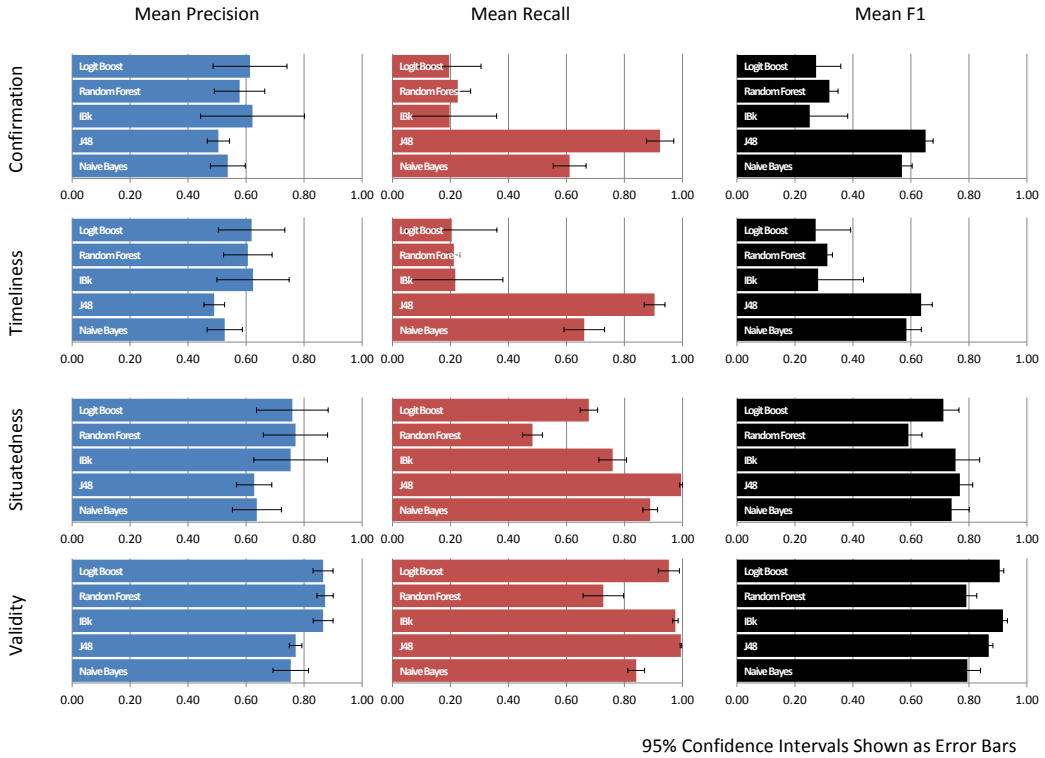

Fig. 5. Leave one out cross validation results comparing classifiers across all 4 classes using the POS feature type. These results represent performance for *unknown* event types. As previously the J48 classifier is best overall.

As expected we see a larger standard deviation than for the 10 fold cross validation experiment, with a maximum F1 standard deviation of 0.15 for IBk. Overall the standard deviation was small however, and the J48 results were stable with a maximum F1 standard deviation of 0.04. When taken across all 4 classes the superiority of the J48 classifier is statistically significant again.

7.4 Experiment 4: TREC Microblog Classification

The fourth experiment tested our approach on tweets from the TREC 2012 microblog dataset. As our 4 class types have not been studied before in this type of social media context we cannot directly compare to any published work. However we show our results using the TREC dataset so that other researchers can benchmark against them in the future; our TREC class labels are available to researchers on request.

We performed a leave one out cross validation experiment again, training using the first 4 event datasets and testing using the fifth TREC 2012 dataset. The raw TREC 2012 dataset contains 16 million tweets sampled between Jan and Fed 2011. We focused on the subset of tweets identified in the TREC cluster MB57 for the Chicago Blizzard of Jan 2011. The experiment conditions were identical to experiment 3, using a POS feature phrase type and J48 decision tree classifier across all 4 classes. The results can be seen in Fig. 6.

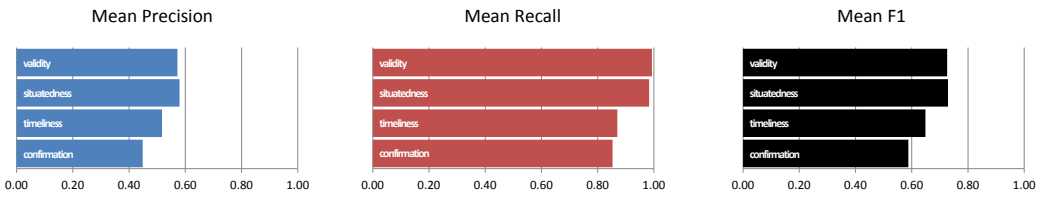

Fig. 6. Benchmark results across all 4 classes testing on the TREC 2012 microblog dataset for the MB57 Chicago blizzard 2011 event. A J48 classifier using a POS feature phrase type was trained on the 4 event dataset, then tested on the TREC dataset

Overall the results in experiment 4 are slightly worse when compared to experiment 3, but the inter-class relative performance is the same. This worse performance is statistically significant for 2 of the 4 classes when compared to experiment 3. We think this is a result of the smaller size of the TREC cluster compared to our other event datasets, and the fact that some class labels (e.g. 'insitu' and 'neg') have little or no examples represented.

8. DISCUSSION

8.1 Discussion of results

We have found from practical experience that our new up-scaled geoparsing approach copes well with the requirement to add new focus areas on demand as breaking news stories develop in real-time. For the Ukrainian crisis case study we started running our system with a basic set of global cities as we hoped something might happen that was news worthy during January. Once breaking news reports came in that Donetsk airport was under attack we quickly added Donetsk and the surrounding regions as new focus areas and were able to geoparse content very quickly. This has given us confidence that our approach will work well when we run full-scale prototype user trials with journalists later in the REVEAL project.

From our experience analysing geosemantics in social media it is clear that one or two phrases in a sentence can really help to indicate how a location is being referred to. The degree of specialization of language used varies between classes. The 'situatedness' class has the most limited feature vocabulary (e.g. 'eyewitness report', 'reporting live', 'journalist at the scene'). The 'timeliness' class has the widest vocabulary with many variants of past/future/present tense used when referencing a location. We think that the superior overall performance of the decision tree classifiers (i.e. J48 and RandomForest) is due to the way key vocabulary maps to single branches on each decision tree, resulting in very specialized and accurate classifiers. Other instance-based classifiers (e.g. IBk) factor in contributions from vectors of many features and can be misled when classifying classes where there is really only a limited vocabulary needed. We found many of the tweets from IBk which were classified incorrectly also had a low classification confidence score, resulting in failure to get above the 0.7 classification probability threshold and a low recall score.

It should be noted that our choice of configuration parameter values (e.g. semantic distance, N value for topN features) is determined empirically. It would be possible to adopt a principled approach to automatically optimize these settings, using techniques such as principle component analysis or information gain metrics.

For the F1 metric we see that the non-boosted J48 decision tree classifier was best for 3 classes (i.e. 'confirmation', 'timeliness', 'situatedness') with an F1 score for known events of between 0.70 to 0.82, and the IBk k-nearest neighbour classifier best for the remaining class (i.e. 'validity') with a score of 0.92. For the precision metric the IBk classifier was best overall with a precision score of between 0.76 and 0.88; however this often was at the expense of a low recall. This classifier might be worth considering if an end user analyst wants to keep false positives to a minimum and does not mind missing some content. An important aspect of our work is demonstrating resilience to unknown event types. In the world of breaking news there are a boundless number of news story types that might be reported on, and maintaining a training corpus for each one would be unrealistic. The J48 classifier has a F1 score of between 0.64 and 0.87 for unknown event types, which is only slightly worse than the results for known events. We therefore consider these results very promising for our use case.

8.2 Limitations of research

One limitation of our approach is the use of supervised learning and the associated need for training data to be compiled for each class (e.g. situatedness). An unsupervised technique would be able adapt to other classes more easily without the need for costly manual data collection and annotation steps. Another limitation with our work is that it does not consider image and video features, relying only on linguistic patterns within text. Whilst more difficult and potentially error prone, approaches that couple text analysis with image and video forensics offer the possibility to do cross-checking of factual data (e.g. location, weather, time of day, faces in shot) related to rumours in addition to relevance pre-filtering of contextual content around rumours.

8.3 Application of research

To put these results into a journalistic context lets imagine during a breaking news story there is a single key eyewitness video on YouTube that people are starting to tweet about which debunks a false claim (i.e. rumour) from one of the news story stakeholders. Even assuming our worst 'situatedness' classification scores (i.e. P 0.63, R 0.99, F1 0.77) for an unknown event we would correctly classify over 60% of situated tweets (i.e. reports insitu such as eyewitness reports) from a coverage of almost 100% of the content available. This would be a problem if we only had access to a single tweet about the eyewitness video as we might filter it out erroneously and miss it. However, important eyewitness videos usually go viral and get retweeted & commented quickly. This means that after the first 3 or 4 tweets about the eyewitness video we would have correctly classified at least one of them, labelled the tweet as 'insitu' eyewitness content of relevance and presented the mentioned YouTube URI to the journalist (e.g. as an eyewitness video rendered on a map). We would have also filtered out the other 95%+ tweets that mentioned relevant locations but were not eyewitness reports, reducing the volume of content the journalist needs to analyse dramatically.

9. CASE STUDIES

9.1 Case Study 1: False Rumours of NYSE Flooding during Hurricane Sandy October 2012

During October 2012 hurricane Sandy battered the south coast of the USA and storm surges flooded large parts of New York. At around 8pm on 30th October a tweet

appeared “BREAKING: Confirmed flooding on NYSE. The trading floor is flooded under more than 3 feet of water.” from Twitter account @Comfortablysmug. It is not clear if this was the original source of this story about the New York Stock Exchange (NYSE) but it was the most influential [Wemple 2012]. Within minutes the story had moved from social media to mass media when CNN forecaster Chad Myers mentioned it during Piers Morgan’s TV program and the Weather Channel tweeted[10] about it.

Unfortunately the story was false. There was a 40 minute period after the initial false rumour tweet from @Comfortablysmug, and more importantly the Weather Channel's retweet of this rumour, when conflicting stories were rife on TV and social media. Eventually both the WeatherChannel and CNN acknowledged on TV and via Twitter that these earlier reports were in fact false.

This section outlines our work examining how geoparsing and geosemantic analysis can, for this NYSE case study, pre-filter content for journalists without losing important key content. Journalists for this type of case study are most interested in spatially grounding content (i.e. the NYSE) and tracing flooding reports back in time to discover their original source. The journalists would ultimately seek to contact the original source and verify the content manually.

We analysed a Twitter dataset crawled during the event in 2012 using flood related filter keywords (i.e. flood, flooding, flooded). We focus on a 5 minute time period from 30th October 2012 01:53:04 to 01:58:17 containing a total of 7,361 tweets. This time period was in the middle of the false news story and represents a time when different sources on TV and social media were reporting the story as both true and false. We geoparsed and geosemantically classified the whole dataset.

For a ground truth we manually identified key tweets from the first 1,000 geoparsed entries (i.e. first 1,000 from the 2,153 New York region filtered tweets) that referenced TV and Twitter reports from either the Weather Channel or CNN (i.e. a total of 114 key tweets from Weather Channel or CNN). Each key tweet was manually labelled as a CONFIRM or DENY. These key tweets would be relevant to any journalist trying to trace retweeted content back in time to find the source of the rumour. We focus on the Weather Channel and CNN since reports from these two mainstream news sources were subsequently used [Wemple 2012] by lots of people to report the rumour as 'verified by a trusted source' and therefore true. The Weather Channel was the first trusted source to report the story as true.

We wanted to see how different types of geosemantic filtering could reduce the volume of tweets whilst retaining intact any ground truth key tweets. We calculated the % volume reduction from the original raw dataset when applying a regional filter (i.e. New York spatial region), location specific filter (i.e. NYSE building) and geosemantic filters for confirmed (i.e. CONFIRM) and not-confirmed classes (i.e. DENY and NA). The results can be seen in Table 2 and a map visualization of the geoparsed dataset can be seen in Fig. 7.

We aggregated the DENY and NA classes since we found that many tweets included claims from CNN in quotes, adding sarcastic comments after the original report (e.g. "NYSE under 3ft of water -- another example of shoddy verification by CNN"). Such sarcastic comments ended up classified as NA not DENY, but this was still a useful result as it differentiated content from the CONFIRM class.

[10] https://twitter.com/weatherchannel/status/263093566065238016

Table 2. Effectiveness of filter types for NYSE rumour case study

	Ground truth tweets retained in filtered dataset		
Filter type	# Tweets (% of raw)	Precision	Recall
Raw data filter: flood keywords	7,361 (100%)	1.00	1.00
Geoparsed filter: NY region	2,153 (29%)	1.00	1.00
Geoparsed filter: NYSE	739 (10%)	1.00	1.00
Geosemantics filter: NYSE filter: CONFIRM	346 (5%)	0.68	1.00
Geosemantics filter: NYSE filter: DENY & NA	339 (5%)	0.47	0.85

The number of ground truth tweets remaining in our filtered dataset is reported via a recall metric. We report the number of correctly classified ground truth tweets via a precision metric. Equation 2 defines these metrics. We are interested in the tradeoff between reduction in dataset size verses loss of key content. As such recall is the most important metric for this case study. It can be seen that even when filtering the dataset down to 5% of its original size we do not lose any confirm tweets and only lost a few deny tweets. This result indicates that geosemantic filtering could be very useful for journalists working on this type of real-world news story.

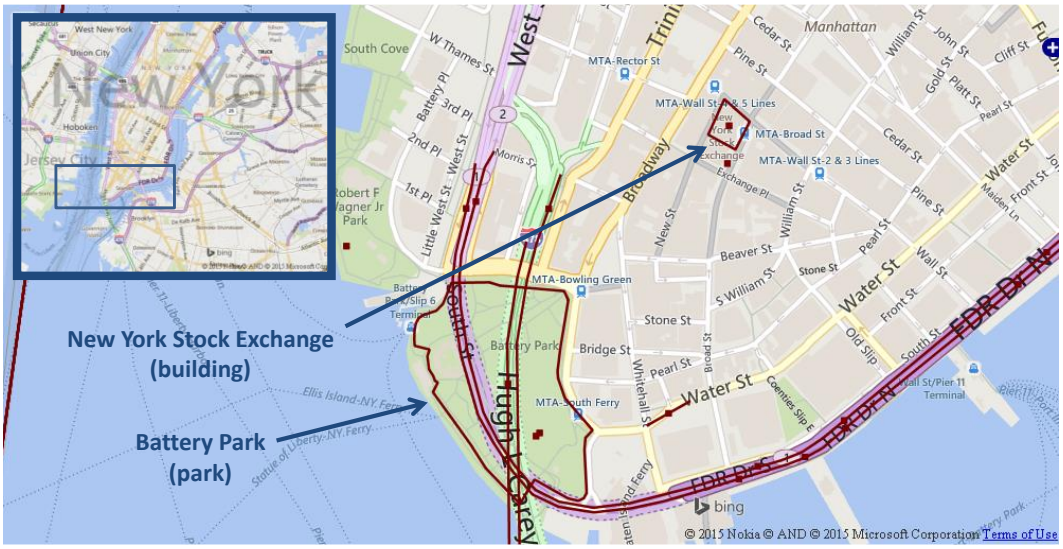

Fig. 7. Screenshot from geospatial visualization of content from Twitter for 30th October 2012 matching flooding keywords. Red lines indicate geoparsed regions, streets and buildings. Clicking on each region, street or building brings up a list of images / videos ranked by mention frequency. Mapping courtesy of Bing Maps.

9.2 Case Study 2: Conflicting Claims between Ukraine and Russia over Who Controlled Donetsk Airport in January 2015

In early January 2015 Ukrainian troops withdrew from Donetsk airport's main terminal after many weeks of bitter fighting with pro-Russian separatists. This was a symbolic victory for the separatists as Donetsk airport had grown in symbolic value

over the months before even though it was now left in ruins. In the days before Ukrainian TV and Russian TV had run conflicting reports claiming their side controlled the airport, with both sides citing unverified evidence from social media to make their case. Which report was true? Both social media and TV channels were alive with conflicting reports and debate over what was really happening. Eventually the truth emerged and the Ukrainian government admitted they has lost control at a cost of 6 dead and 16 wounded soldiers.

This section outlines our work examining how geoparsing can spatially ground content relating to known news events. Journalists for this type of case study are interested in compiling sets of eyewitness videos from social media content that is within the time window of the event and spatially nearby the event location. Journalists will typically look to correlate geographic features between videos to verify content location, try to identify combatants and their nationalities and cross-check videos to identify any inconsistencies as part of the journalist's manual verification process.

We analysed a Twitter, YouTube and Instagram dataset crawled during the event in 2015 using conflict type keywords in English, Russian and Ukrainian. We focus on a 24 hour time period from the 20th January 2015 containing a total of 332,000 content items. This time period was just after the Dontesk airport had fallen and both sides were claiming victory. We geoparsed all content items and produced a content map clustered by location. We then selected the Donetsk airport location and retrieved its clustered set of You Tube URI's ranked by URI mention frequency. This spatially grounded ranked set of URI's represents the type of information our system can provide journalists to spatially ground and filter large volumes of content. The map of clustered content for Donetsk airport can be seen in Fig. 3.

For a ground truth we used two news reports that appeared that day from Russian TV broadcaster Life News[11]. These news reports represent a verified account of what happened that day and cited 4 key You Tube videos as evidence that pro-Russian separatists had won the battle and taken Ukrainian soldiers prisoner as a result. We found that 3 of the 4 ground truth You Tube videos appeared in our Donetsk airport cluster, ranked at positions 10, 14 and 28 out of the top 30. The news reports and You Tube videos can be seen in Fig. 8. This shows that our geoparsing approach is able to spatially ground and rank relevant social media content for a real-world news story without too much information loss. It also shows we can display this clustered and ranked content to journalists in a way that can assist their existing manual verification procedures.

[11] http://lifenews.ru/

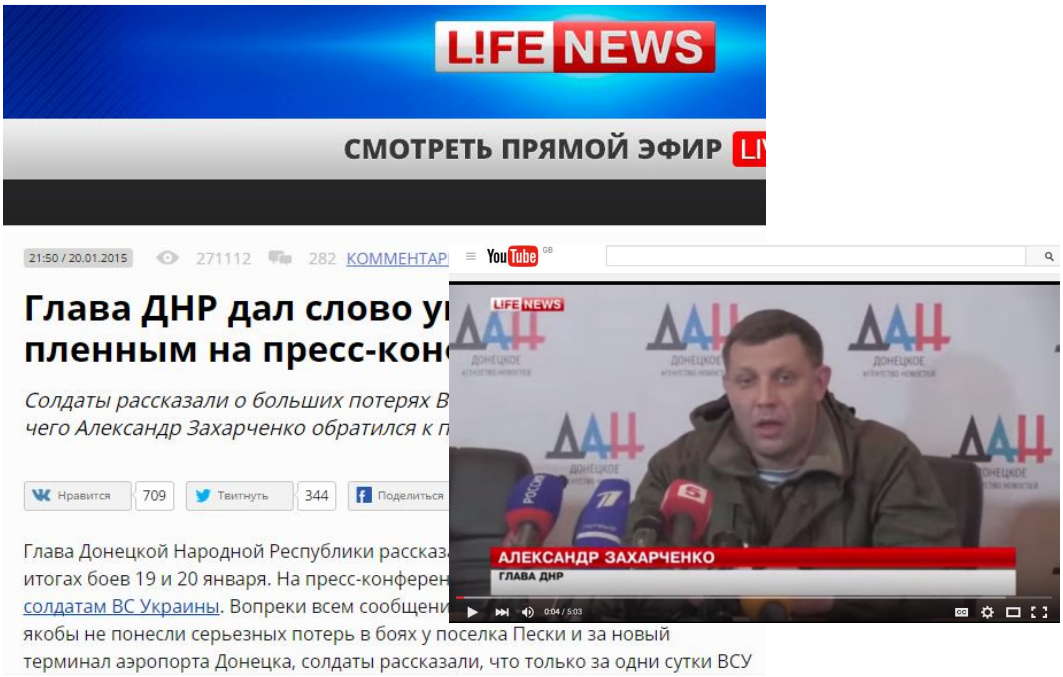

Fig. 8. Fall of Dontsk airport 20th January 2015 in pictures. Images include one of the YouTube videos that appear in both our Donetsk Airport location cluster and the ground truth Life News reports that feature these videos. Source: Life News and You Tube

10. CONCLUSIONS

When a news story is breaking journalists are under a lot of time pressure to be first to publish and this makes verification of content associated with rumours a difficult task. Journalists often need to trace content relating to rumours back in time to identify the source and then contact that source manually. They also need to spatially correlate content relating to event locations to allow cross-checking of eyewitness reports and verify often biased accounts of what happened. We have shown in this paper that both scalable geoparsing approaches and geosemantic classification of content can help journalists manage large volumes of social media content and spatio-temporally ground it for manual analysis and verification.

We have described a scalable geoparsing approach that can handle on-demand requests for focus areas during real-time breaking news stories. The use of Apache Storm means that geoparsing processes can be run in parallel over a computing cluster to handle multiple news stories and focus areas. In our Ukraine crisis case study we show that we can spatially ground large volumes of social media content and geospatially visualize it so journalists can easily find spatially relevant content from nearby locations in a specific time window.

We have described and evaluated a novel geosemantic feature extraction approach able to classify content in terms of 'confirmation', 'timeliness' , 'situatedness' and 'validity'. Whilst not perfect, with classification F1 scores of 0.7 to 0.82, we have had informal positive feedback from our Journalist end users with regards our pre-filtering performance. The classifier is also effective on new unseen event types which is very important since news stories cover a wide range of topics that would be impossible to pre-define in advance. In our NYSE case study we show that geosemantic pre-filtering can reduce raw content volumes by 95% without losing much content relevant to the rumour under investigation. Reducing content volumes

by 95% means journalists can spend more time verifying key content and less time sifting through irrelevant content.

Our two case studies represent exemplars of core tasks Journalists face in day to day news room verification of user generated content. The first use case represents a typical strategy of tracing the propagation of a rumour back in time by looking at retweets and comments on the original post which might no longer be visible (e.g. if it had been deleted). This is needed as a Journalist will usually try to contact the original post's author directly (e.g. via a phone call) as part of verifying the source of the rumour. The second use case represents another strategy of finding contextual content which allows cross-checking of suspicious posts, especially those with eyewitness images or videos. Journalists typically use tools such as Geofeedia to map content and find images which are spatially and temporally nearby. They then look for things within these contextual images that can be cross-checked with the original image, such as faces of passers-by, buildings in the background, number plates or street signs. If evidence from contextual content cross-checks correctly with the image or video under investigation then it will be considered more credible.

We have had positive feedback from journalists in the REVEAL project with regards the application of both geoparsing and geosemantics to help filter breaking news content. The current tools journalists use for discovery and verification tasks such as TweetDeck (i.e. tracking tweets from multiple people), Google Reverse Image Search (i.e. finding similar images) and TinEye (i.e. checking image metadata and similarity) are effective but time consuming to use. This means journalists can only verify a limited amount of content before they hit their publication deadlines. Being able to reduce the volume of raw content, without losing key content needed to debunk rumours, would allow journalists to focus their verification effort more efficiently. This in turn could improve rumour verification and lead to less mistakes being made under the time pressures associated with breaking news.

In the context of advancement in trust and veracity research this work represents a step forward towards more scalable approaches to both the analysis of content veracity and the analysis of trustworthiness in the sources propagating rumours. Our work is limited to supporting domains where there is an existing manual verification process, such as journalism or intelligence analysis. We empower the human analyst to scale up the volume of content they are able to consider when verifying rumours through better filtering of irrelevant content and better identification of contextual spatio-temporal content for cross-checking of facts. Considering larger volumes of content should improve the accuracy of each individual analyst's decision making, whilst not compromising on timescales, something which is a key challenge when working on verifying breaking news.

For next steps we are planning to conduct further experiments on sets of social media content relating to multi-lingual news events based in languages such as Russian, Italian and German. The aim is to examine how resilient multi-lingual geosemantic feature extraction can be when compared to results for English language content. We are also planning a set of ethnographic studies in the REVEAL project to look at the subjective judgements made in the news room when selecting or rejecting user generated content as evidence for breaking news stories. We hope these studies will provide insights that allow us to develop a trust and credibility model to support a semi-automated interactive process where journalists explore different views on large volumes of evidence derived from social media content items. We expect our

work on geoparsing and geosemantics will provide important features that this model can use to infer new facts relating to trust and credibility.

ACKNOWLEDGMENTS

The authors would like to thank journalists at Deutsche Welle for their valuable guidance helping us understand the complexities of the real-world journalistic verification process for user generated content.

REFERENCES

Alan Ritter, Sam Clark, Mausam and Oren Etzioni, 2011. Named entity recognition in tweets: An experimental study. In *Procedings of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK

Alexandre Davis, Adriano Veloso, Altigran S. da Silva, Wagner Meira Jr., Alberto H. F. Laender, 2012. Named entity disambiguation in streaming data. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 815 - 824, Jeju, Republic of Korea

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, 2002. Thumbs up? Sentiment Classification using machine learning techniques. In *Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing (Philadelphia, PA, 2002)*. Association for Computational Linguistics, Morristown, NJ, 79 - 86

Claire Grover, Richard Tobin, Beatrice Alex and Kate Byrne, 2010. Edinburgh-LTG: TempEval-2 System Description. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010*. ACL 2010, Uppsala, Sweden, 333 - 336

Craig Silverman (Ed.), 2013. Verification Handbook. *European Journalism Centre*

Craig Silverman, 2015. Lies, Damn Lies, and Viral Content. How News Websites Spread (and Debunk) Online Rumors, Unverified Claims, And Misinformation. *Tow Center for Digital Journalism*, Columbia Journalism School

Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Yiannis Kompatsiaris, 2015. The CERTH-UNITN Participation @ Verifying Multimedia Use 2015. In *MediaEval Benchmarking Initiative for Multimedia Evaluation* 2015 (*MediaEval-2015)*, Wurzen, Germany

Eliot Higgins, 2014. A Beginner's Guide to Geolocating Videos. *Bellingcat*

Emanuele Bastianelli, Danilo Croce, Roberto Basili, and Daniele Nardi, 2013. UNITOR-HMM-TK: Structured Kernel-based learning for Spatial Role Labeling. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. ACL 2013, Atlanta, Georgia, 573 - 579

Erik Wemple, 2012. Hurricane Sandy: NYSE NOT flooded!, *The Washington Post (October 2012)*

Hanan Samet, Jagan Sankaranarayanan, Michael D. Lieberman, Marco D. Adelfio, Brendan C. Fruin, Jack M. Lotkowski, Daniele Panozzo, Jon Sperling and Benjamin E. Teitler, 2014. Reading News with Maps by Exploiting Spatial Synonyms. *Communications of the ACM*, vol.57, no.10, 64-77

Hector Llorens, Estela Saquete and Borja Navarro, 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010*. ACL 2010, Uppsala, Sweden, 284 - 29

Hong Yu and Vasileios Hatzivassiloglou, 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (2003)*.Stroudsburg, PA, USA, 129-136

Ian H. Witten and Eibe Frank, 2011. Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann Publishers*

Ian Soboroff, Iadh Ounis, Craig Macdonald and Jimmy Lin, 2012. Overview of the TREC2012 Microblog Track. In *Proceedings of the 20th Text REtrieval Conference (TREC 2012)*. Gaithersburg MD, USA

Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, Christopher Collins, 2014. *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12

Jochen Spangenberg and Nicolaus Heise, 2014. News from the Crowd: Grassroots and Collaborative Journalism in the Digital Age. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW 2014)*. Seoul, Korea, 765-768

Joshua Lieberman and Chris Goad, 2008. Geosemantic Web Standards for the Spatial Information Infrastructure. In *Creating Spatial Information Infrastructures*, Peter van Oosterom and Sisi Zlatanova (Eds.) CRC Press, 119-128

Judith Gelernter and Nikolai Mushegian, 2011. Geo-parsing Messages from Microtext. *Transactions in GIS*, Vol 15, Issue 6, 753–773

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard and Niraj Aswani, 2013. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, 83-90

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede, 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* Vol. 37, No. 2, 267–307

Marc Verhagen, Roser Saur, Tommaso Caselli and James Pustejovsky, 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden, 57 - 62

María Teresa Vicente-Díez, Julián Moreno Schneider and Paloma Martínez, 2010. UC3M system: Determining the Extent, Type and Value of Time Expressions in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010.* ACL 2010, Uppsala, Sweden, 329 - 332

Marieke van Erp, Giuseppe Rizzo and Raphaël Troncy, 2013. Learning with the Web: Spotting Named Entities on the intersection of NERD and Machine Learning. In *Proceedings of the 3rd Workshop on Making Sense of Microposts (#MSM2013)*. Rio de Janeiro, Brazil

Minqing Hu and Bing Liu, 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (SIGKDD-2004).* Seattle, Washington, 168–177.

Oren Tsur, Dmitry Davidov and Ari Rappoport, 2010. A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*. Washington, DC

Parisa Kordjamshidi, Steven Bethard, Marie-Francine Moens, 2012. SemEval-2012 task 3: spatial role labeling. In *Proceeding of the 6th International Workshop on Semantic Evaluation (SemEval '12*), 365-373

Peter D. Turney, 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics*. Philadelphia, 417 - 424

Raz Schwartz, Mor Naaman and Rannie Teodoro, 2015. Editorial Algorithms: Using Social Media to Discover and Report Local News. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media (ICWSM-15)*. Oxford, UK

Ronen Feldman, 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, Volume 56 Issue 4, April 2013 , 82-89

Samantha Finn, Panagiotis T. Metaxas, Eni Mustafaraj, Megan O'Keefe, L. Tang, S. Tang, Laura Zeng, 2014. TRAILS: A System for Monitoring the Propagation of Rumors on Twitter. Computation and Journalism Symposium, NYC, NY, 2014

Samuel Carton, Souneil Park, Nicole Zeffer, Eytan Adar, Qiaozhu Mei and Paul Resnick, 2015. Audience Analysis for Competing Memes in Social Media. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media (ICWSM-15)*. Oxford, UK

Steven Bird, Ewan Klein, and Edward Loper, 2009. Natural Language Processing with Python—Analyzing Text with the Natural Language Toolkit, *O'Reilly Media*

Stuart E. Middleton, 2015. Extracting Attributed Verification and Debunking Reports from Social Media: MediaEval-2015 Trust and Credibility Analysis of Image and Video, In *MediaEval Benchmarking Initiative for Multimedia Evaluation* 2015 (*MediaEval-2015)*, Wurzen, Germany

Stuart E. Middleton, Lee Middleton and Stefano Modafferi, 2014. Real-Time Crisis Mapping of Natural Disasters Using Social Media. *Intelligent Systems*, IEEE, vol.29, no.2, 9-17

Zhe Zhao, Paul Resnick and Qiaozhu Mei, 2015. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. In *Proceedings of the 24th International Conference on World Wide Web ((IW3C2)*, Florence, Italy

Received March 2015; revised August 2015; revised Oct 2015; accepted xxxx