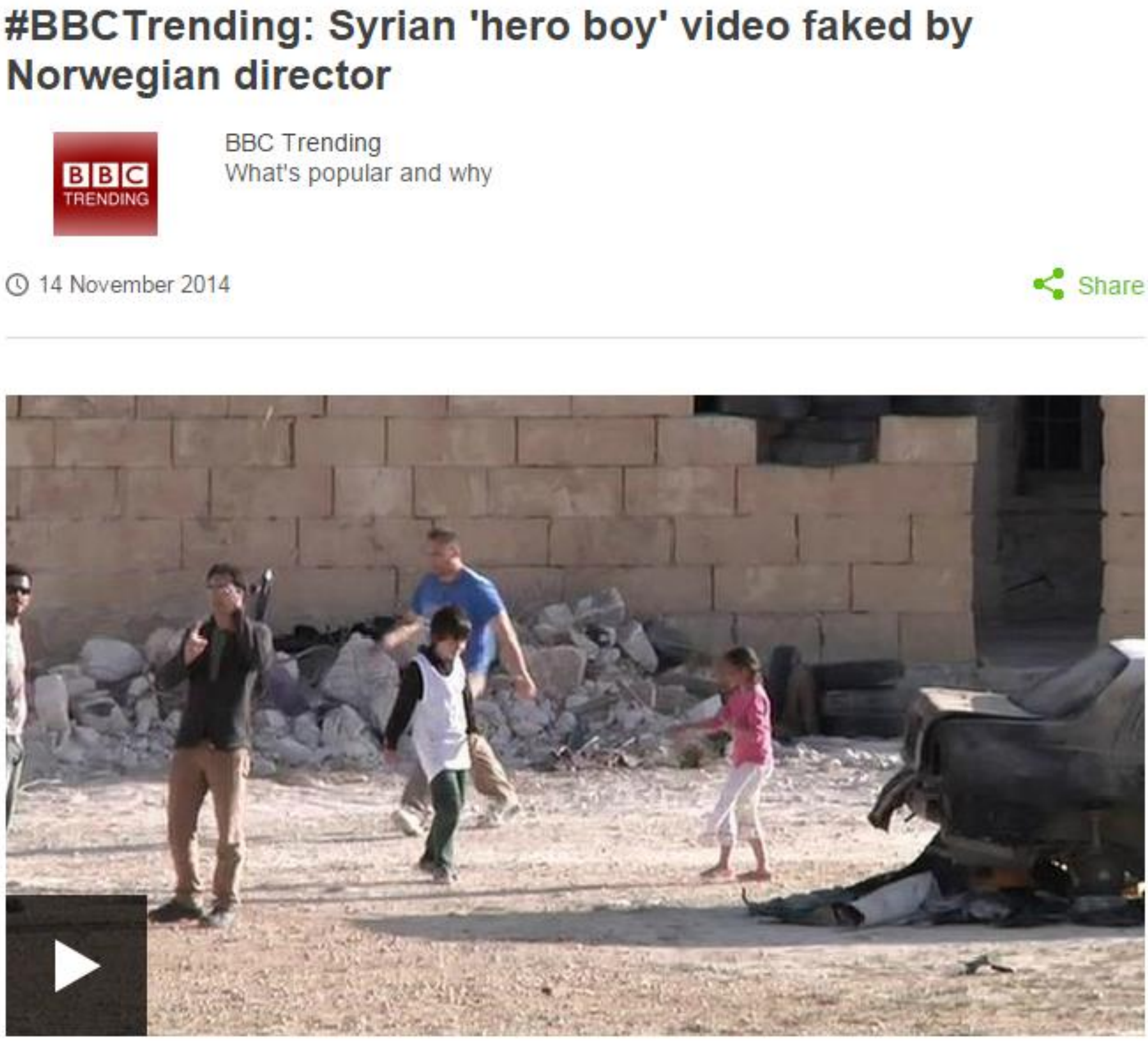
29th Sept 2015

Semi-Automated Extraction of Attributed Verification and Debunking Reports from Social Media

Content from social media sites such as Twitter, YouTube, Facebook and Instagram are becoming an important part of modern journalism. Of particular importance to real-time breaking news is amateur on the spot incident reports and eyewitness images and videos. With breaking news having tight reporting deadlines, measured in minutes not days, the need to quickly verify suspicious content is paramount [1] [3]. Journalists are increasingly looking to pre-filter and automate the simpler parts of the verification process.

Current tools available to journalists can be broadly categorized as dashboard and in-depth analytic tools. Dashboard tools display filtered traffic volumes, trending hashtags and maps of content by topic, author and/or location. In-depth analysis tools use techniques such as sentiment analysis, social network graph visualization and topic tracking. These tools help journalists manage social media content but unverified rumours and fake news stories on social media are becoming both increasingly common [2] and increasingly difficult to spot. The current best practice for journalistic user generated content (UGC) verification [1] 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 ways to automate the simpler verification steps, empowering journalists and helping them to focus on cross-checking tasks that most need human expertise. We are creating a trust and credibility model able to process real-time evidence extracted using a combination of natural language processing, image analysis, social network analysis and semantic analysis. This article describes our work on text analysis, extracting and processing fake and genuine claims from tweets referencing suspicious images and videos. Our central hypothesis is that the 'wisdom of the crowd' is not really wisdom at all when it comes to verifying suspicious images and videos. Instead it is better to rank evidence from Twitter according to the most trusted and credible sources in a way similar to human journalists. We describe a semi-automated approach, automatically extracting claims about real or fake content and their source attributions and comparing them to a manually created list of trusted sources. A cross-checking step ranks conflicting claims and selects the most trustworthy evidence on which to base a final fake/real decision.

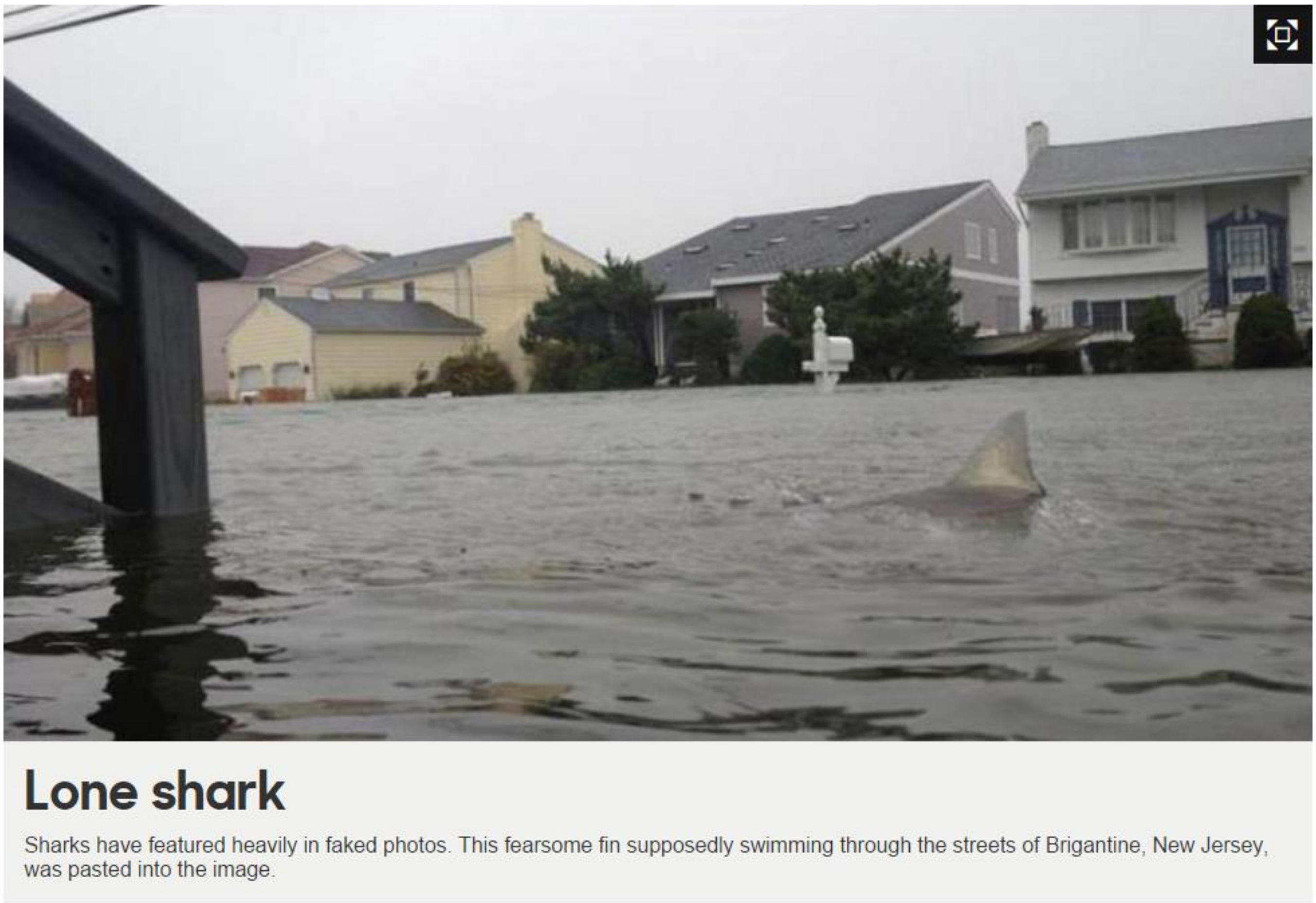


#BBCTrending: Syrian 'hero boy' video faked by Norwegian director

BBC Trending
What's popular and why

14 November 2014

Figure 1: BBC debunking report for a viral video about a Syrian 'hero boy' - Source BBCNews



Lone shark

Sharks have featured heavily in faked photos. This fearsome fin supposedly swimming through the streets of Brigantine, New Jersey, was pasted into the image.

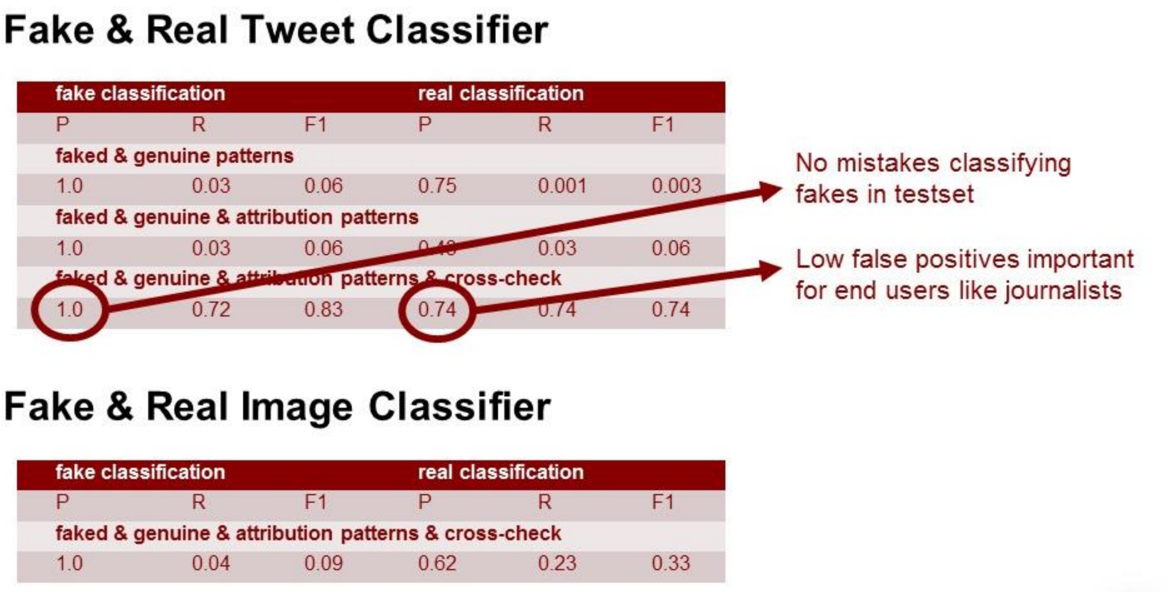
Figure 2: Photoshopped shark in a viral image during hurricane Sandy 2012 - Source BBCNews

MediaEval-2015 Verification Challenge

The MediaEval 2015 Verifying Multimedia Use challenge [4] [6] is an annual event which tests international teams of computer science researchers on their ability to verify multimedia content. Teams receive sets of tweets mentioning suspicious images or videos and must use multimedia features and textual patterns to decide it its real or fake. A fake is considered to be a manipulated image (e.g. photoshopped images) or an original image presented in the wrong context (e.g. photos of the wrong warzone presented as an atrocity). See figures 1 and 2 for examples. All teams submit their image and video classifications, which are then compared by the challenge organizers to a hidden ground truth based on a human assessment of the content. The teams are scored by how many classifications they get right, ranked and a winner chosen. The winning team will have the best balance between a low error rate and a high classification rate.

Results

Our approach's strength is that it has a very low false positive rate, and in fact made no mistakes at all when classifying the MediaEval-2015 Verifying Multimedia Use challenge dataset. Figure 3 highlights the low false positive rate with a maximum precision score of 1.0. Full details can be found in the working notes paper [5].



Fake & Real Tweet Classifier

fake classification			real classification		
P	R	F1	P	R	F1
faked & genuine patterns					
1.0	0.03	0.06	0.75	0.001	0.003
faked & genuine & attribution patterns					
1.0	0.03	0.06	0.43	0.03	0.06
faked & genuine & attribution patterns & cross-check					
1.0	0.72	0.83	0.74	0.74	0.74

No mistakes classifying fakes in testset

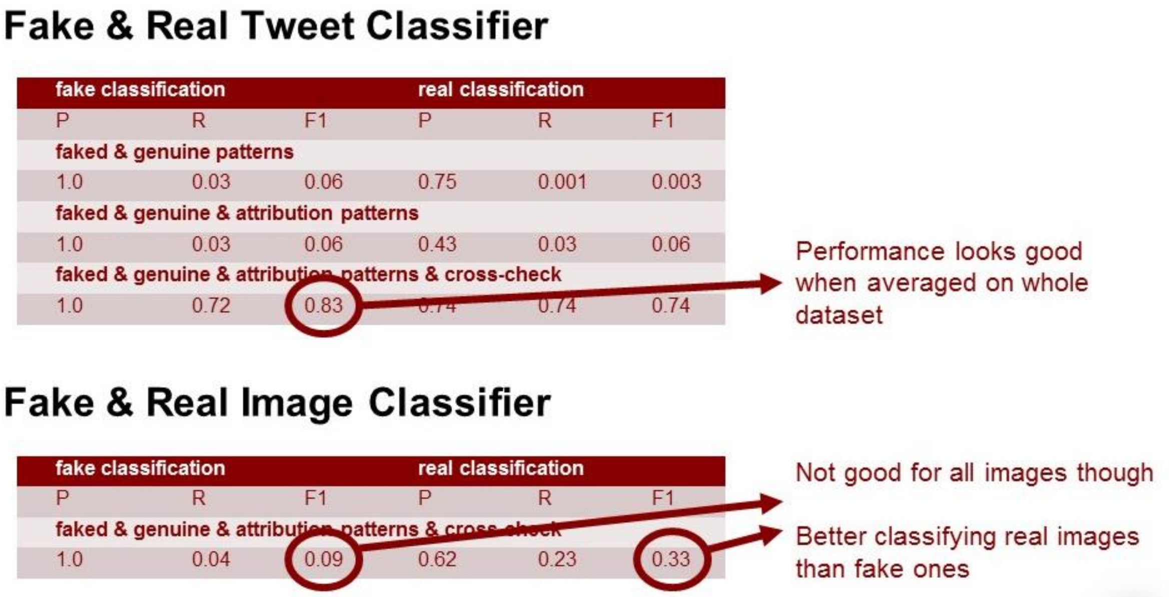
Low false positives important for end users like journalists

Fake & Real Image Classifier

fake classification			real classification		
P	R	F1	P	R	F1
faked & genuine & attribution patterns & cross-check					
1.0	0.04	0.09	0.62	0.23	0.33

Figure 3: Results from MediaEval-2015 Verifying Multimedia Use challenge (1)

Our approach's weakness is a lower classification rate, since not all images have tweeted claims about its verification or debunking status and as such were not always able to reach a decision and had to label the content as 'unknown'. Figure 4 highlights the low classification rate with a modest recall score.



Fake & Real Tweet Classifier

fake classification			real classification		
P	R	F1	P	R	F1
faked & genuine patterns					
1.0	0.03	0.06	0.75	0.001	0.003
faked & genuine & attribution patterns					
1.0	0.03	0.06	0.43	0.03	0.06
faked & genuine & attribution patterns & cross-check					
1.0	0.72	0.83	0.74	0.74	0.74

Performance looks good when averaged on whole dataset

Fake & Real Image Classifier

fake classification			real classification		
P	R	F1	P	R	F1
faked & genuine & attribution patterns & cross-check					
1.0	0.04	0.09	0.62	0.23	0.33

Not good for all images though

Better classifying real images than fake ones

Figure 4: Results from MediaEval-2015 Verifying Multimedia Use challenge (2)

Technology

We extract from textual patterns in tweets claims about the image being fake or real, and attribution statements about the source of the content. We compare attributed source named entities (e.g. BBC News) to a list of trusted sources in the same way a human journalist might so. Our trust and credibility model is based on a classic natural language processing pipeline involving tokenization, Parts of Speech (POS) tagging, named entity recognition and relational extraction. Full technical details can be seen in the working notes paper [5].

Future

In the context of journalistic verification these results are promising. Given enough tweeted claims about an image or video we can rank the most trustworthy and provide a highly accurate classification result. This means that once images and videos, such as eyewitness content, go viral on twitter we will be able to provide a real-time view on their verification status. Our approach does not replace manual verification techniques - someone still needs to actually verify the content - but it can rapidly alert journalists to trustworthy reports of verification and/or debunking. This in turn should speed up the verification cycle and allow the 'time to publish' to be shortened.

We are working on a range of trust and verification algorithms in addition to this work. Examples include automated image verification via a cross-check of known facts, automatically downloading the historical weather and time of day for an event and checking this against image features (e.g. if it’s raining in an image but it was a dry day for the event the image must be a fake). We are also developing interactive analytical visualizations to both display clusters of content geolocated on maps, and display temporally sampled content on timelines. These visualizations will allow journalists to explore contextual social media content, quickly finding evidence that can be used for cross-checking facts about a story.

We hope to release a live demonstration tool in the Spring of 2016. Announcements will be made via the REVEAL website. Alternatively you can follow us on Twitter to be the first to know!

Acknowledgement

The work presented in this article is part of the research and development in the REVEAL project (grant agreement 610928), supported by the 7th Framework Program of the European Commission.

About the author

Stuart E. Middleton is a senior research engineer at the University of Southampton IT Innovation Centre. His main research interests are social media, sensor systems, data fusion and semantics. Stuart has a PhD in Computer Science from the University of Southampton.

@stuart_e_middle   @IT_Innov

http://www.it-innovation.soton.ac.uk http://users.ecs.soton.ac.uk/sem/

REVEAL project, @RevealEU

http://revealproject.eu/

References

[1] Silverman, C. (Ed.), 2013. Verification Handbook. European Journalism Centre

[2] Silverman, C. 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

[3] Spangenberg, J. Heise, N. 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

[4] Boididou, C. Andreadou, K. Papadopoulos, S. Dang-Nguyen, D. Boato, G. Riegler, M. Kompatsiaris, Y. 2015. Verifying Multimedia Use at MediaEval 2015. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany

[5] Middleton, S.E."Extracting Attributed Verification and Debunking Reports from Social Media: MediaEval-2015 Trust and Credibility Analysis of Image and Video", MediaEval-2015, Wurzen, Germany, Sept 2015

[6] MediaEval-2015, http://wwwu.edu.uni-klu.ac.at/miriegle/mediaeval/index2015.html