Extracting Attributed Verification and Debunking Reports from Social Media: MediaEval-2015 Trust and Credibility Analysis of Image and Video

Stuart E. Middleton
University of Southampton IT Innovation Centre
Southampton, UK
sem@it-innovation.soton.ac.uk

ABSTRACT

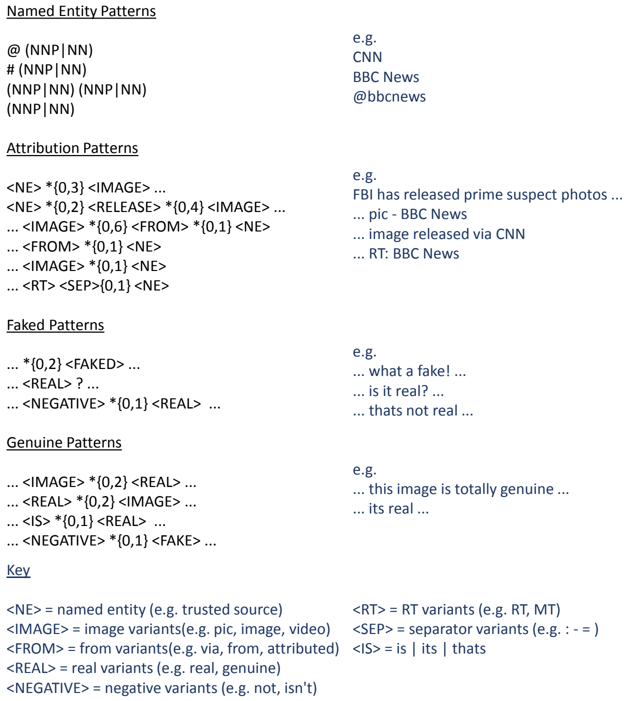
Journalists are increasingly turning to technology for pre-filtering and automation of the simpler parts of the verification process. We present results from our semi-automated approach to trust and credibility analysis of tweets referencing suspicious images and videos. We use natural language processing to extract evidence from tweets in the form of fake & genuine claims attributed to trusted and untrusted sources. Results for team UoS-ITI in the MediaEval 2015 Verifying Multimedia Use task are reported. Our 'fake' tweet classifier precision scores range from 0.94 to 1.0 (recall 0.43 to 0.72), and our 'real' tweet classifier precision scores range from 0.74 to 0.78 (recall 0.51 to 0.74). Image classification precision scores range from 0.62 to 1.0 (recall 0.04 to 0.23). Our approach can automatically alert journalists in real-time to trustworthy claims verifying or debunking viral images or videos.

1. INTRODUCTION

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 [5] [7]. 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 [6] and increasingly difficult to spot. The current best practice for journalistic user generated content (UGC) verification [5] 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 paper 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.



Named Entity Patterns

@ (NNP|NN)
# (NNP|NN)
(NNP|NN) (NNP|NN)
(NNP|NN)

e.g.
CNN
BBC News
@bbcnews

Attribution Patterns

<NE> *{0,3} <IMAGE> ...
<NE> *{0,2} <RELEASE> *{0,4} <IMAGE> ...
... <IMAGE> *{0,6} <FROM> *{0,1} <NE>
... <FROM> *{0,1} <NE>
... <IMAGE> *{0,1} <NE>
... <RT> <SEP>{0,1} <NE>

e.g.
FBI has released prime suspect photos ...
... pic - BBC News
... image released via CNN
... RT: BBC News

Faked Patterns

... *{0,2} <FAKED> ...
... <REAL> ? ...
... <NEGATIVE> *{0,1} <REAL> ...

e.g.
... what a fake! ...
... is it real? ...
... thats not real ...

Genuine Patterns

... <IMAGE> *{0,2} <REAL> ...
... <REAL> *{0,2} <IMAGE> ...
... <IS> *{0,1} <REAL> ...
... <NEGATIVE> *{0,1} <FAKE> ...

e.g.
... this image is totally genuine ...
... its real ...

Key

<NE> = named entity (e.g. trusted source)
<IMAGE> = image variants(e.g. pic, image, video)
<FROM> = from variants(e.g. via, from, attributed)
<REAL> = real variants (e.g. real, genuine)
<NEGATIVE> = negative variants (e.g. not, isn't)

<RT> = RT variants (e.g. RT, MT)
<SEP> = separator variants (e.g. : - =)
<IS> = is | its | thats

Figure 1: Verification Linguistic Patterns. These patterns are encoded as regex patterns matching on both phrases in content and their associated POS tags (e.g. NN = noun, NNP = proper noun).

2. APPROACH

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. The innovation in our approach lies with our choice of regex patterns, which are modelled on how journalists verify fake and genuine claims by looking at the source attribution for each claim. This allows us to provide a novel conflict resolution approach based on ranking claims in order of trustworthiness. We use the Python NLTK toolkit [1], weak stemming, Punkt sentence tokenizer and Treetagger POS tagger. To extract fake and genuine

MediaEval-2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany.

claims we use a set of regex patterns (see Figure 1) matching both terms and POS tags. To discover attribution we use a combination of named entity matching and regex patterns.

Our semi-automated approach to named entity matching is based on a list of a priori known trusted and untrusted sources. We can either learn an entity list automatically using information theoretic weightings (i.e. TF-IDF) or create a list manually (i.e. using a journalists trusted source list). All news providers have long lists of trusted sources for different regions around the world so this information is readily available. For the MediaEval 2015 Verifying Multimedia Use task we created a list of candidate named entities by first running the regex patterns on the dataset. We then manually checked each entity via Google search (e.g. looking at Twitter profile pages). We removed any named entities which we considered a journalist would not have in a list of trusted or untrusted sources. We kept news organizations, respected journalists and well cited bloggers and experts. Creating these lists took under two hours (570 named entities checked, 60 accepted).

We chose these regex patterns based on the frequency of text patterns for source attribution, fake and genuine claims in the MediaEval-2015 devset. Other researchers have published linguistic patterns used to detect rumours [3] [8] [4] but our combination of fake/genuine claims and source attribution is novel, using insights from the well-established journalistic verification processes for User Generated Content (UGC).

We assign a confidence value to each matched pattern based on its source trustworthiness level. Evidence from trusted authors is more trusted than evidence attributed to trusted authors, which is more trusted than other unattributed evidence. In a cross-check step we choose the most trustworthy claims to use for each image URI. If there is evidence for both a fake and genuine claim with an equal confidence we assume it is fake (i.e. any doubt = fake).

Table 1: Fake and Real Tweet Classification for Devset

fake classification			real classification		
P	R	F1	P	R	F1
faked & genuine patterns					
0.89	0.007	0.01	1.0	0.0007	0.001
faked & genuine & attribution patterns					
0.89	0.007	0.01	0.99	0.05	0.11
faked & genuine & attribution patterns & cross-check					
0.94	0.43	0.59	0.78	0.51	0.61

Table 2: Fake and Real Image Classification for Devset

fake classification			real classification		
P	R	F1	P	R	F1
faked & genuine & attribution patterns & cross-check					
0.96	0.10	0.19	0.95	0.19	0.32

3. RESULTS

The MediaEval 2015 Verifying Multimedia Use task is to classify tweets about images and videos as real, fake or unknown. Details of the task datasets, ground truth and evaluation methodology used can be found in [2]. Results in Table 1 & Table 2 show fake and real classification performance for the devset, with Table 3 & Table 4 showing the testset. Journalists ultimately want to find verified genuine content that they can use in breaking news stories. As such whilst the MediaEval-2015 Verifying Multimedia Use task is focussed on classifying fake content we also report results for the harder problem of classifying real content. We report image classification accuracy as well as classification accuracy of tweets referring to these images.

Our first fully automated run used the 'faked & genuine' regex patterns applied to each tweet independently without lists of trusted sources. The second semi-automated run used in addition the source attribution regex patterns, matching attributed named entities to a manually created list of trusted and untrusted sources. The final semi-automated run added the cross-check step, making a decision not on the basis of each tweet alone but rather using the most trustworthy evidence available after cross-checking all tweets referring to a specific image or video. This final approach is the most realistic one for our journalistic use case; eyewitness images and videos going viral during a breaking news story will typically have hundreds of comments on Twitter before journalists discover them and attempt verification.

Table 3: Fake and Real Tweet Classification for Testset.

fake classification			real classification		
P	R	F1	P	R	F1
faked & genuine patterns (run-1)					
1.0	0.03	0.06	0.75	0.001	0.003
faked & genuine & attribution patterns (run-3)					
1.0	0.03	0.06	0.43	0.03	0.06
faked & genuine & attribution patterns & cross-check (run-4)					
1.0	0.72	0.83	0.74	0.74	0.74

Table 4: Fake and Real Image Classification for Testset

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

4. CONCLUSION

When it comes to verifying claims about suspicious images and videos our hypothesis is that the 'wisdom of the crowd' is not really wisdom at all and it is better to rank evidence from Twitter in order of the most trusted and credible sources. We have developed a semi-automated trust and credibility model based on this intuition and well known journalistic verification principles.

When applied to classifying tweets in isolation, our approach has a high precision and low recall, making it of limited value. When we cross-check tweets, ranking by trustworthiness and picking only the most trusted claims our approach is much more useful, with a high precision (0.94+) and average recall (0.43+). The ultimate goal of course is to classify images as fake (including use of image in the wrong context) or real not just the tweets that refer to them. Our classifier was able to classify 4-10% of fake images, getting it right 96-100% of the time. For the harder problem of classifying real images our approach was able to classify 19-23% of images, getting it right 62-95% of the time.

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.

5. ACKNOWLEDGEMENTS

This work is part of the research and development in the REVEAL project (grant agreement 610928), supported by the 7th Framework Program of the European Commission. The authors would like to thank journalists at Deutsche Welle for their valuable insights into the journalistic verification process.

6. REFERENCES

[1] Bird, S. Klein, E. Loper, E. 2009. Natural Language Processing with Python—Analyzing Text with the Natural Language Toolkit, *O'Reilly Media*

[2] 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

[3] Boididou, C. Papadopoulos, S. Kompatsiaris, Y. Schifferes, S. Newman, N. 2014. Challenges of computational verification in social multimedia. *In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland

[4] Carton, S. Park, S. Zeffer, N. Adar, E. Mei, Q. Resnick, P. 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

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

[6] 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

[7] 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

[8] Zhao, Z. Resnick, P. Mei, Q. 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