Social Media Verification

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

sem@it-innovation.soton.ac.uk @stuart_e_middle @IT_Innov @RevealEU

www.it-innovation.soton.ac.uk

UoS-ITI Team

Overview

- Problem Statement
- Approach
- Results
- Discussion
- Suggestions for Verification Challenge 2016

Problem Statement

Verification of Images and Videos for Breaking News

- Breaking News Timescales
  - Minutes not hours - its old news after a couple of hours
  - Journalists need to verify copy and get it published before their rivals do
- Journalistic Manual Verification Procedures for User Generated Content (UGC)
  - Check content provenance - original post? location? timestamp? similar posts? website? ...
  - Check author / source - attributed or author? known (un)reliable? popular? reputation? post history? ...
  - Check content credibility - right image metadata? right location? right people? right weather? ...
  - Phone the author up - triangulate facts, quiz author to check genuine, get authorization to publish
- Automate the Simpler Verification Steps
  - Empowering journalists
  - Increases the volume of contextual content that can be considered
  - Focus humans on the more complex & subjective cross-checking tasks
    - Contact content authors via phone and ask them difficult questions
    - Does human behaviour 'look right' in a video?
    - Cross-reference buildings / landmarks in image backgrounds to Google StreetView / image databases
    - ... see the VerificationHandbook » http://verificationhandbook.com/

Approach

Attribute evidence to trusted or untrusted sources

- Hypothesis
  - The 'wisdom of the crowd' is not really wisdom at all when it comes to verifying suspicious content
  - It is better to rank evidence according to the most trusted & credible sources like journalists do
- Semi-automated approach
  - Manually create a list of trusted sources
  - Tweets » NLP » Extract fake & genuine claims & attribution to sources » Evidence
  - Evidence » Cross-check all content for image / video » Fake/real decision based on best evidence
- Trustworthiness hierarchy for tweeted claims about images & videos
  - Claim = statement that its a fake image / video or its genuine
  - Claim authored by trusted source ☑ ☑ ☑
  - Claim authored by untrusted source ☒ ☒ ☒
  - Claim attributed to trusted source ☑ ☑
  - Claim attributed to untrusted source ☒ ☒
  - Unattributed claim ☑

Approach

Regex patterns

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

Results

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

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

Results

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

Results

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

Discussion

Application to our journalism use case

- Classifying tweets in isolation (fake and real) is of limited value
  - High precision (89%+) but low recall (1%)
- Cross-check tweets then ranking by trustworthiness
  - No false positives for fake classification using testset
  - High precision (94%+) with average recall (43%+) looking across events in devset and testset
  - Typically viral images & videos will have 100's of tweets before journalists become aware of them so a recall of 20% is probably OK in this context
- Image classifiers
  - Fake image classifier » High precision (96-100%) but low recall (4-10%)
  - Real image classifier » High precision (62-95%) but low recall (19-23%)
- Classification explained in ways journalists understand & therefore trust
  - Image X claimed verified by Tweet Y attributing to trusted entity Z
  - We can alert journalists to trustworthy reports of verification and/or debunking
- Our approach does not replace manual verification techniques
  - Someone still needs to actually verify the content!

Suggestions for Verification Challenge 2016

Focus on image classification not Tweet classification

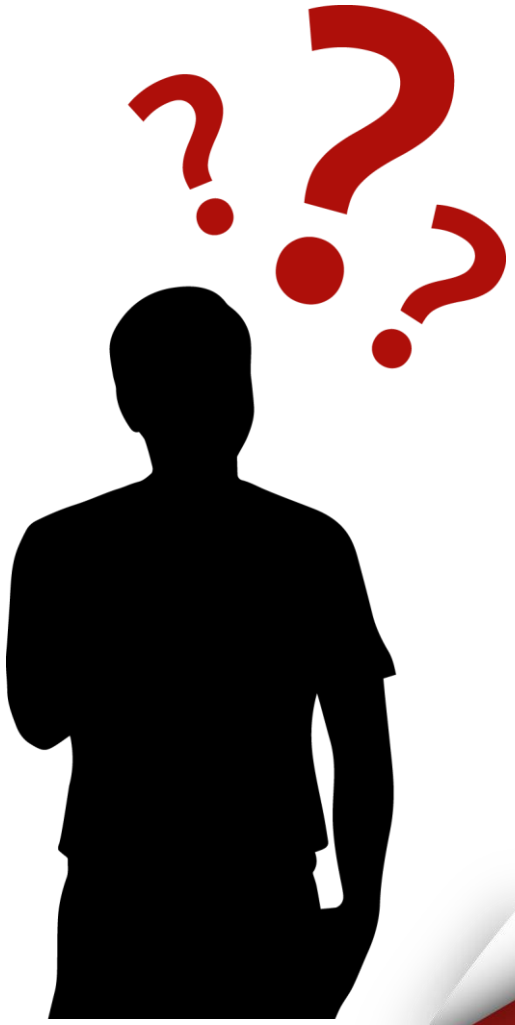
- The long term aim is to classify the images & videos NOT the tweets about them
  - **Suggestion** » Score image classification results as well as tweet classification results
- End users usually wants to know if its real, not if its fake
  - Classifying something as fake is usually a means to an end (e.g. to allow filtering)
  - **Suggestion** » Score results for fake classification & real classification

Improve the Tweet datasets to avoid bias to a single event

- **Suggest** using leave one event out cross validation when computing P/R/F1
- **Suggest** removing tweet repetition
  - Some events (e.g. Syrian Boy) contain many duplicate tweets with a different author
  - A classifier might only work well on 1 or 2 text styles BUT score highly as they are repeated a lot
- **Suggest** evenly balancing number of tweets per event type to avoid bias
  - Devset - Hurricane Sandy event has about 84% of the tweets
  - Testset - Syrian Boy event has about 47% of the tweets

Many thanks for your attention!

Any questions?



Stuart E. Middleton
University of Southampton IT Innovation Centre

email: sem@it-innovation.soton.ac.uk
web: www.it-innovation.soton.ac.uk
twitter:@stuart_e_middle, @IT_Innov, @RevealEU