Social Computing for Verifying Social Media Content in Breaking News

Stuart E. Middleton
University of Southampton

Symeon Papadopoulos
Centre for Research and Technology Hellas

Yiannis Kompatsiaris
Centre for Research and Technology Hellas

Social media is the place to go for both journalists and the general public when news events break, offering a real-time source of eyewitness images and videos through platforms like YouTube, Instagram, and Periscope. Yet, the value of such content as a means of documenting and disseminating breaking news is compromised by the increasing amount of content misuse and false claims in social media. To this end, cost-effective social-computing solutions for user-generated content verification are crucial for retaining the value and trust in social media for breaking news.

Most people have a smartphone in their pocket today, so eyewitnesses experiencing an event like a terror attack will often post real-time claims, such as the numbers dead or injured in a location, to Twitter or Facebook. Eyewitness images and videos will also be uploaded to sites like YouTube and Instagram or even streamed live to sites like Periscope. For events such as the 2015 Paris shootings,[1] the first eyewitness videos of the various shootings were posted within 5 to 10 minutes of the event happening. This was followed about 20 to 30 minutes later with verified news reports from sources such as *Le Figaro*, the BBC, and CNN. In other cases, verifying eyewitness or user-generated media and claims can take much longer, from hours to even days, as, for instance, in the case of Malaysia Airlines Flight 17, which was shot down on 17 July 2014.

In many cases, as soon as a breaking news event starts trending on Twitter, it is accompanied by considerable numbers of false claims and content misuse.[2] This involves the use of multimedia for misinforming the public and misrepresenting people, organizations, and events. Misuse practices range from publishing content that has been digitally tampered using photo-editing software to falsely associating content with an unfolding event. The paper "Detection and Resolution of Rumours in Social Media: A Survey"[2] contains an extensive discussion on the problem of rumor detection in social media.

Published by the IEEE Computer Society
1089-7801/18/$33.00 USD ©2018 IEEE

Given the grave societal and economic impact of having misused content and false claims featured in mainstream news, it becomes extremely important for news organizations to be able to verify eyewitness media in very short time. To this end, journalists are turning to social-computing approaches to automatically analyze and verify[3] user-generated content in real time. The eventual hope is that cost-effective social computing can reduce the time spent on verification to time scales nearer to real time.

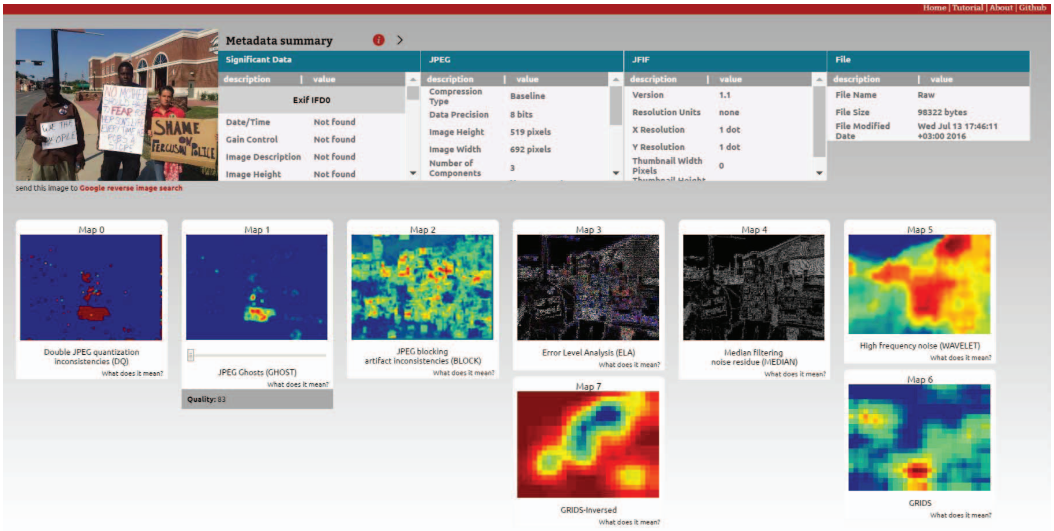
SOCIAL-MULTIMEDIA FORENSICS AND SUPERVISED VERIFICATION

Methods from the field of digital forensics are often used for assessing the veracity of multimedia items (images or videos) posted online. Some methods focus on the analysis of information encoded in the metadata of multimedia content, such as EXIF (Exchangeable Image File) information, which is often associated with JPEG and TIFF images. Several types of digital manipulation, such as the use of photo-editing software, leave traces in the form of metadata unless special care is taken to remove them, and analysis of these traces can detect manipulations. Unfortunately, several of the most popular social media platforms, including Facebook and Twitter, automatically remove much of the metadata from posted content, rendering metadata-based methods useless for content obtained from these platforms.

Other forensics-based methods aim at uncovering traces of manipulation in the visual content itself. In images, such methods[4] can detect cases of splicing and copy–move operations—for example, inpainting of a part of one image into a second, or replication of a part of an image within the same image. Methods can leverage the uniqueness of noise patterns introduced by the capturing device in order to detect whether an image contains traces from another image captured by a different device. Other methods focus on patterns associated with the color filter arrays of modern image-capturing equipment. Splice detection methods exploit traces left by the JPEG compression process, working on the basis that the splicing of two different images and the subsequent recompression will leave detectable traces in the final JPEG file.

While all of the above methods yield satisfactory results when applied on well-controlled test samples, they have been found to exhibit poor performance in real cases.[5] One of the reasons that state-of-the-art methods fail to detect manipulations in media content published on the web is that such content is often the result of numerous intermediate operations, including resizing, cropping, and recompression, which have an obfuscating effect on the traces of digital manipulation. Examples are Twitter and Facebook, both of which automatically resize and recompress all images uploaded to them.[6]

Recent work in the FP7 REVEAL project (see Figure 1) addresses the poor performance of individual tampering-detection methods[7] by generating tampering probability heat maps based on a number of complementary forensic-analysis algorithms. The inclusion of multiple image-forensics algorithms and side-by-side comparisons gives a powerful means to journalists to understand where possible digital tampering has occurred. The problem of identifying digital manipulations in video content is even more challenging compared to the case of images, and it is further exacerbated in cases where such content is sourced from video-sharing and social-networking platforms such as YouTube and Facebook. The Horizon 2020 progam's InVID project is looking into resilient approaches for video forensics building upon the TUNGSTENE commercial forensics engine.


Home | Tutorial | About | Github
send this image to Google reverse image search
Map 0
Double JPEG quantization inconsistencies (DQ)
What does it mean?
Map 1
JPEG Ghosts (GHOST)
What does it mean?
Quality: 83
Map 2
JPEG blocking artifact inconsistencies (BLOCK)
What does it mean?
Map 3
Error Level Analysis (ELA)
What does it mean?
Map 4
Median filtering noise residue (MEDIAN)
What does it mean?
Map 5
High frequency noise (WAVELET)
What does it mean?
Map 7
GRIDS-Inversed
What does it mean?
Map 6
GRIDS
What does it mean?

Metadata summary

Significant Data

description	value
Exif IFD0	
Date/Time	Not found
Gain Control	Not found
Image Description	Not found
Image Height	Not found

JPEG

description	value
Compression Type	Baseline
Data Precision	8 bits
Image Height	519 pixels
Image Width	692 pixels
Number of Components	3

JFIF

description	value
Version	1.1
Resolution Units	none
X Resolution	1 dot
Y Resolution	1 dot
Thumbnail Width Pixels	0

File

description	value
File Name	Raw
File Size	98322 bytes
File Modified Date	Wed Jul 13 17:46:11 +03:00 2016

Figure 1. A digital-forensics platform for image verification (screenshot taken from the Media Verification Assistant).

Some cases of content misuse are not detectable by using forensics analysis—for example, when an image from a past event is reposted as being associated with an unfolding event. These require other methods that attempt to detect misuse by analyzing contextual cues from social media sources.[8] A typical approach adopted by such methods is to extract a variety of trust-oriented features from social media posts, and the accounts generating these posts, and to use them for training machine-learning models based on historically labeled cases of fake and real posts. Such methods have been shown to yield very high fake-post-detection accuracy.[9]

EYEWITNESS MEDIA AND FACT EXTRACTION

The standard workflow for automated fact checking[10] involves the monitoring of data sources, fact identification, fact extraction, and fact checking. The challenges for social computing mostly involve fact identification and extraction. Once a fact is extracted, it can be checked either manually or automatically against databases from sites such as PolitiFact, FactCheck.org, Snopes, and Wikipedia.

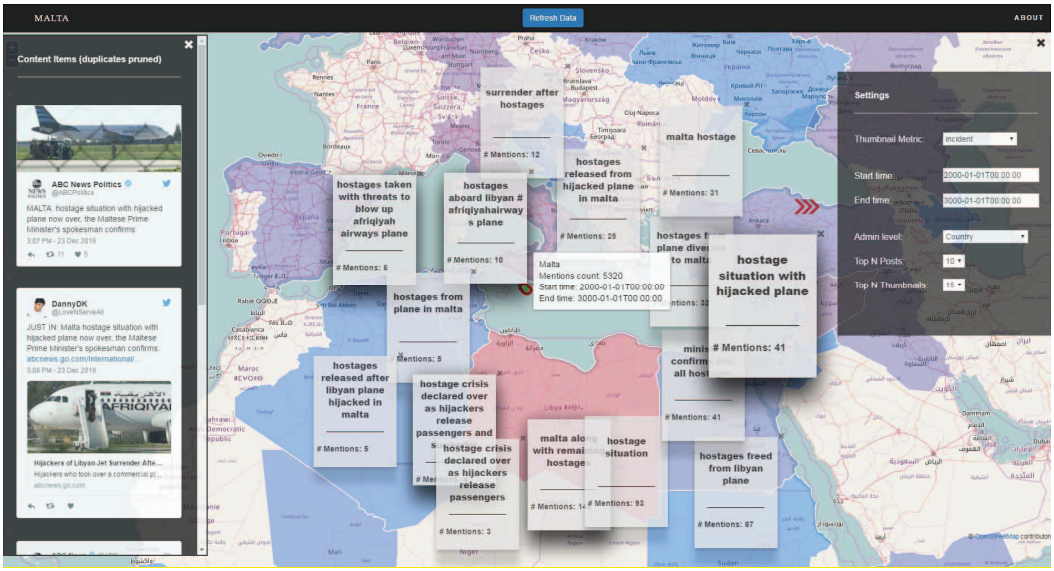
Factual claims come in many forms. The most important for social computing are factual assertions, contextual statements associated with a fact, and contextual statements associated with the trustworthiness of the fact. Factual assertions themselves can be true, false, half-truths, or exaggerations. Contextual statements can allow a true representation of a fact or misrepresent it by suggestions of a false location, actor, or time stamp. Contextual text can also introduce ideological cues and loaded language to bias the interpretation of the fact. Finally, contextual statements can suggest trustworthiness, such as attribution to a trusted source or claims of previous verification that themselves might be subject to falsehoods or deliberate bias.

Fact identification approaches, especially for news-related sources, try to classify sentences into nonfactual, unimportant factual, and "check worthy" factual statements[11] so that they can be filtered prior to fact extraction. Fact extraction is a type of information extraction problem that runs alongside information extraction techniques for concepts such as event, topic, location, and time. In the past, approaches such as argumentative zoning[12] were applied successfully to extract factual statements on well-structured and trustworthy scientific documents. However, the text in web and social media sources is often neither well structured nor trustworthy, so new approaches are being explored.

Early work in this area focused on verb phrase patterns (e.g., "was elected to") to extract facts via systems such as OLLIE (Open Language Learning for Information Extraction).[13] These used part-of-speech (POS) tagging, dependency parsing, and distant supervision coupled with seed attributes and bootstrapping to provide unsupervised fact extraction. In particular, they were able to capture the “long tail” of factual statements, which is very important for the contextual interpretation (e.g., “Putin made a deal with the separatists”). Later advances,[14] motivated by the need to answer queries in search engines, added noun phrase patterns (e.g. “Obama’s wife”) very successfully. Typically, such approaches exploit large databases of attribute names and noun phrases such as Freebase and DBpedia.

Automated fact checkers use either domain-specific databases (e.g., PolitiFact) or web-scale datasets (e.g., DBpedia). Recently, there has been a trend of real-time crowdsourcing of fact checking during events such as US political rallies, with the Trump–Clinton presidential debates being the latest example. Fake news sites have also been increasing in number and can easily mislead readers[15] into trusting misinformation based on a credible but false source attribution. The iCheck system[16] is a good example where domain-specific heuristics extract fact types that are visualized via a crowdsourcing interface for users to check claims and up- or down-vote them.

Work from the REVEAL project[17] has taken these ideas one step further to help journalists verify breaking news. Automated fact extraction using semantic grammars, seeded with linguistic phrases originating from journalists, is used to extract evidence from social media content about news events such as incident reports, facts about damage, the numbers of the dead and injured, people, locations, and attributed sources. User-generated content from the scene of a breaking event, not yet syndicated via news organizations, is particularly important for journalists. Supervised-learning algorithms are employed within REVEAL to identify and filter posts containing eyewitness images and videos. This type of social computing is coupled with real-time visualizations (see Figure 2), allowing journalist to quickly find contextual content such as original mentions of claims for subsequent verification.


MALTA
Refresh Data
ABOUT
surrender after hostages
# Mentions: 12
malta hostage
# Mentions: 31
hostages released from hijacked plane in malta
# Mentions: 25
hostages taken with threats to blow up afriqiyah airways plane
# Mentions: 6
hostages aboard libyan # afriqiyahairways plane
# Mentions: 10
hostages from plane diverted to malta
# Mentions: 32
hostage situation with hijacked plane
# Mentions: 41
hostages from plane in malta
# Mentions: 5
minis confirm all host
# Mentions: 41
hostages released after libyan plane hijacked in malta
# Mentions: 5
hostage crisis declared over as hijackers release passengers and s
# Mentions: 4
hostage crisis declared over as hijackers release passengers
# Mentions: 3
malta alo with remai hostage
# Mentions: 14
hostage situation
# Mentions: 93
hostages freed from libyan plane
# Mentions: 87

Content Items (duplicates pruned)

ABC News Politics @ABCPolitics
MALTA: hostage situation with hijacked plane now over, the Maltese Prime Minister's spokesman confirms:
3:07 PM - 23 Dec 2016
11 5

DannyDK @LoveNServeAll
JUST IN: Malta hostage situation with hijacked plane now over, the Maltese Prime Minister's spokesman confirms: abcnews.go.com/International/...
3:08 PM - 23 Dec 2016
Hijackers of Libyan Jet Surrender Afte...
Hijackers who took over a commercial pl...
abcnews.go.com

Malta
Mentions count: 5320
Start time: 2000-01-01T00:00:00
End time: 3000-01-01T00:00:00

Settings
Thumbnail Metric: incident
Start time: 2000-01-01T00:00:00
End time: 3000-01-01T00:00:00
Admin level: Country
Top N Posts: 10
Top N Thumbnails: 15

Figure 2. An interactive real-time visualization mapping extracted facts and eyewitness media in posts about the December 2016 Malta plane hijacking (screenshot taken from the Journalist Decision Support System).

CONCLUSION

We highlighted in this article the potential of employing social-computing approaches for speeding up the task of verifying user-contributed information and content sourced from social media platforms. The problem is complex and calls for a variety of approaches, each targeting different challenges stemming from the characteristics of user-generated content, including high volume, inconsistent quality, and a lack of provenance information. Multimedia forensics targets the actual content of multimedia. Supervised verification is best suited to cases where contextual features can be extracted and labeled training sets of fake and real examples are available. Fact extraction and visualization approaches target text-based sources that contain references to different elements of an event, such as people, times, and locations.

The REVEAL project is one of the first efforts to bring together those technologies under a single platform that could provide comprehensive verification support to professional users; details on the successful user evaluation of pilot prototypes can be found in “D7.2 User Evaluation Report.”[18] However, there is still a long way to go before such tools are widely used by newsrooms and journalists day to day. Figure 3 provides an overview of the projects and datasets useful to researchers interested in automating verification tasks for social media and news-related content.

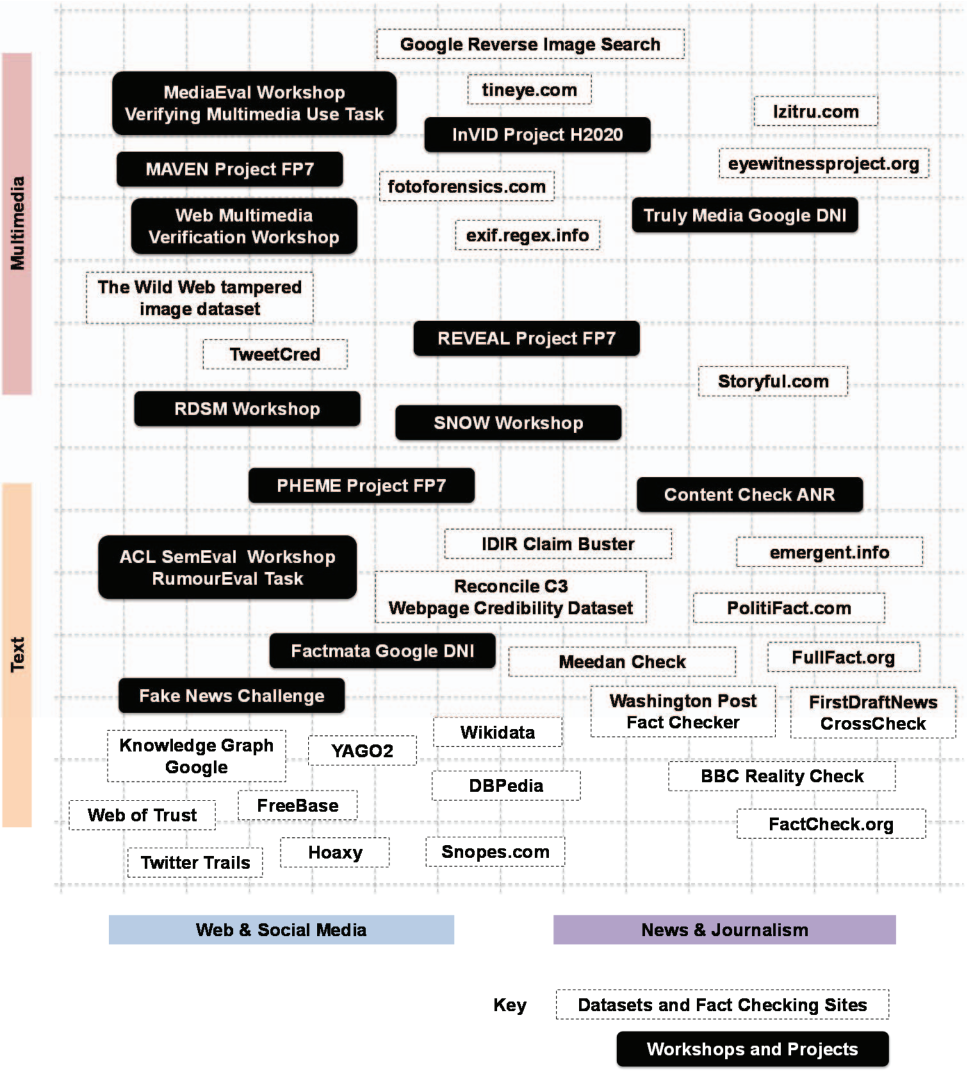


Figure 3. The research and dataset landscape for researchers interested in verification of social media and news-related content.

One key challenge involved in delivering such an integrated solution is the lack of an appropriate human–computer interaction (HCI) approach that would empower users (e.g., journalists) to make optimal use of the technologies described above. Given the extensive use of algorithms, an effective HCI approach would need to build the trust of users by providing intuitive control and a clear explanation of the results. Ultimately, users are in charge of the whole process and will make the final decision with respect to whether a piece of user-generated content should be considered authentic or not. Moreover, support for collaborative work among teams of journalists is another key social-computing challenge that is missing from existing news provider in-house solutions, which instead employ general-purpose communication and messaging platforms such as Slack and WhatsApp.

In conclusion, the problem of real-time verification of user-generated content is expected to remain unsolved in the near future, but marked improvements have already been achieved on individual parts of the verification process thanks to social-computing approaches incorporating intelligent information processing. In the future, we anticipate considerable progress on this problem by incorporating the latest advances from deep learning. One example would be employing generative adversarial networks[19] to build highly accurate and robust models for visually distinguishing between tampered-with and untampered-with regions in multimedia content. Another example would be novel HCI approaches focusing on the explainability of automatically generated results and the collaborative aspects of the verification process.

ACKNOWLEDGMENTS

This work was supported by the FP7 REVEAL and Horizon 2020 InVID projects, funded partially by the EC under grant agreements 610928 and 687786, respectively.

REFERENCES

1. S. Wiegand and S.E. Middleton, “Veracity and Velocity of Social Media Content during Breaking News: Analysis of November 2015 Paris Shootings,” *Third Workshop on Social News on the Web, Companion of the 25th International World Wide Web Conference* (SNOW 16), 2016.
2. A. Zubiaga et al., *Detection and Resolution of Rumours in Social Media: a Survey* arXiv preprint, 2017; doi.org/arXiv:1704.00656.
3. C. Silverman, *Verification Handbook: a Definitive Guide to Verifying Digital Content for Emergency Coverage*, European Journalism Centre, 2014.
4. G.K. Birajdar and V.H. Mankar, “Digital Image Forgery Detection using Passive Techniques: A Survey,” *Digital Investigation*, vol. 10, no. 3, 2013, pp. 226–245.
5. M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, “Detecting Image Splicing in the Wild (WEB),” *IEEE International Conference on Multimedia & Expo Workshops* (ICMEW 15), 2015, pp. 1–6.
6. M. Zampoglou et al., “Web and Social Media Image Forensics for News Professionals,” *Tenth International AAAI Conference on Web and Social Media, Social Media in the Newsroom Workshop*, 2016.
7. M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, “Large-scale Evaluation of Splicing Localization Algorithms for Web Images,” *Multimedia Tools Appl.*, vol. 76, no. 4, 2017, pp. 4801–4834.
8. C. Castillo, M. Mendoza, and B. Poblete, “Information Credibility on Twitter,” *Proceedings of the 20th international ACM conference on World Wide Web* (WWW 11), 2011, pp. 675–684.
9. C. Boididou et al., “Learning to Detect Misleading Content on Twitter,” *Proceedings of the International Conference on Multimedia Retrieval*, 2017.
10. M. Babakar and W. Moy, “The State of Automated Factchecking,” *Full Fact*, 2016.
11. N. Hassan et al., “The Quest to Automate Fact-Checking,” *Computation+ Journalism Symposium*, 2015.
12. S. Teufel, *Argumentative Zoning: Information Extraction from Scientific Text*, dissertation PhD diss., University of Edinburgh, 2000.
13. M.S. Mausam et al., “Open Language Learning for Information Extraction,” *Proceedings of Empirical Methods in Natural Language Processing*, 2012.

14. M. Yahya et al., “ReNoun: Fact Extraction for Nominal Attributes,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP 14), 2014, pp. 325–335.
15. W.L. Bennett, *News: The Politics of Illusion*, University Of Chicago Press, 2016.
16. B. Walenz et al., “Finding, Monitoring, and Checking Claims Computationally Based on Structured Data,” *Computation + Journalism Symposium, Columbia University*, Columbia University, 2014.
17. S.E. Middleton and V. Krivcovs, “Geoparsing and Geosemantics for Social Media: Spatio-Temporal Grounding of Content Propagating Rumours to support Trust and Veracity Analysis during Breaking News,” *ACM Transactions on Information Systems*, vol. 34, no. 3, 2016, p. Article 16.
18. P.B. Brandtzaeg et al., *D7.2 User Evaluation Report* REVEAL project, 2017.
19. I. Goodfellow et al., “Generative Adversarial Nets,” *Advances in Neural Information Processing Systems 27* (NIPS), 2014.

ABOUT THE AUTHORS

Stuart E. Middleton is a senior research engineer at the University of Southampton IT Innovation Centre. His main research interests are social media, machine learning, information retrieval, and semantics. Middleton received a PhD in computer science from the University of Southampton. He’s a senior member of ACM. Contact him at sem@it-innovation.soton.ac.uk.

Symeon Papadopoulos is a postdoctoral research fellow at the Information Technologies Institute (ITI), part of the Centre for Research and Technology Hellas (CERTH). His main research interests are multimedia processing and indexing, information retrieval, machine learning, and social network analysis. Papadopoulos received a PhD in computer science from the Aristotle University of Thessaloniki. Contact him at papadop@iti.gr.

Yiannis Kompatsiaris is a senior researcher at the Information Technologies Institute (ITI), part of the Centre for Research and Technology Hellas (CERTH). His main research interests are semantic multimedia analysis, indexing and retrieval, social media and big data analysis, knowledge structures, reasoning, and personalization. Kompatsiaris received a PhD in computer science from the Aristotle University of Thessaloniki. Contact him at ikom@iti.gr.