**Dr. Nikos Sarris** has been working in ATC as a Senior IT Consultant since 2004 and is currently the Head of Operations of the ATC Innovation Lab. He received his PhD from the Aristotle University of Thessaloniki and his Master of Engineering degree in Computer Systems Engineering from the University of Manchester Institute of Science and Technology. Nikos has been working since 1996 in R&D projects as a researcher, project manager and coordinator of large multinational consortia. In the latest years he has mainly been involved in media related projects focusing on the semantic analysis and 'understanding' of multimedia news content. He is the co-editor of a book and has authored numerous scientific publications for international journals and conferences.

**Dr. Eva Jaho** is working as a project manager in the ATC Innovation Lab. She received her PhD and MsC in Networking from the Department of Informatics and Telecommunications of the National & Kapodistrian University of Athens, Greece, in 2011 and 2007 respectively. She received the Diploma degree from the same university department in 2005. Eva has participated in several European research projects in the field of networking and telecommunications and has authored numerous publications for international journals and conferences. Her main research interests lie in the analysis of content networks and data dissemination, as well as social networking applications.

**Type of presentation**: In-use contribution

## Big data analysis techniques for the verification of social media content

The objective is to extract the hidden value from the massive amounts of user-generated content available on the web and create applications that extract maximum utility from content in social media. This includes the development of applications for content validation, as well as search and filtering capabilities based on user preferences.

The changes brought by new web social media technologies mainly refer to the appearance of new types of content providers and new types of content, often referred to as 'new media'. New media is giving the power of speech to the citizens who can now very easily report, blog and send short text messages (e.g., tweets), and rapidly create new content in huge amounts. Traditionally, in the area of news media, conventional journalism has been the main trend, operating with standard news collection and broadcasting procedures while mediating mainstream types of content (e.g., politics, sport, economy, culture, health) from authoritative sources. However, in the last few years, new Internet web technologies have appeared and have disrupted this business process. Traditional news media are getting more and more overcome by the rise of web news services.

The evolution of the Internet is associated to an increasing level of "social involvement" of the users. Social Media (social networking platforms, blogs, etc.) clearly demonstrate the Internet utilization with the highest user involvement. Due to this increasing level of user involvement, social media have become a valuable tool to share information. They offer the opportunity to generate, collect and communicate new knowledge and information in diverse aspects of human life.

SocialSensor[1] is a 3-year FP7 European Integrated Project (2011-2014) aiming to improve the management of information in social media. The tools developed will collect information from social media, and perform the following tasks: (a) *Verification*: ensure that the content posted in social networks is accurate; (b) *Filtering*: filter the content of social media according to individual needs and interests; (c) *Sensing*: discover public opinions and trends; (d) *Analysis*: analyze multimedia User Generated Content (UGC) from the social web; (e) *Visualisation*: present search results in an attractive, easy to understand manner; (f) *Cross-platform operability*: enable searches across different social media platforms and integrate results on a single platform; (g) *Speed*: quick and efficient processing, with adequate accuracy; (h) *Usability*: intuitive and easy to use tools and interfaces.

As social media tools get more sophisticated, and content is being created in massive amounts, there is a need to effectively rate its quality and validity. This entails the analysis of information flows in real or near-real time and the estimation of attributes of users (e.g. trustworthiness, confidentiality), items and metadata (e.g. relevance), as well as the examination of implicit relations among users, content and metadata. Research for content verification is oriented towards the support of users in tracking the provenance/origin of shared media content, estimating the capture location of media content and detecting manipulations and misleading content, through the development of automatic and semi-automatic multimedia verification tools designed for social media content.

An innovative feature conceived in the framework of SocialSensor is a "truth meter" or "alethiometer" (from the Greek word 'αλήθεια', which means truth), which analyses the validity of each tweet or author, based on a 3 'C's framework: Content, Contributor and Context analysis.

***Contributor*** – This involves all data relevant to the source of information, such as its history, its connections and interactions in the social circle and any other information that can assist in the profiling of any contributor of content.

***Content*** – This includes all multimedia analysis methods that can provide clues about the meaning of the content, but also indicate possible manipulations and fraudulent use.

***Context*** – This includes all contextual co-occurrences, which strengthen or weaken the confidence built around several concepts like credibility and suspicious behaviour.

The concept of Alethiometer will be extended in the framework of the REVEAL[2] FP7 project. In this context more research work will be conducted and oriented through the fundamental concepts, i.e., Contributor, Content and Context. The "alethiometer" will also feature advanced image recognition algorithms, enabling to discover if an image is new or has been duplicated, as well as detect the origin of an image and correlate it to the text inputs, for measuring context validity.

Apart from the content itself, the analysis of social media in REVEAL focuses on structural and contextual aspects. Structural aspects pertain to the analysis of links between content contributors and the content itself, and the communities formed. Community detection research can help reveal the structure of the social media network, common interest groups, as well as the most influential users and the most critical discussion topics. It can also be

[1] www.socialsensor.eu
[2] http://www.revealproject.eu

used for modelling the dissemination speed of news in social media, or providing content recommendations to users.

Content analysis entails multimedia indexing, computational stylometry, similarity analysis and general study of relations between content items, and content credibility evaluation. Advanced multimedia indexing is necessary for the detection of named entities and concepts, which is a prerequisite for the more ambitious goal of detecting similarities between media content items. Computational stylometry can help reveal content provenance by finding similarities in writing styles. Similarity analysis is also related to plagiarism, which is linked to credibility evaluation.

Context-centric interpretation involves event recognition, location analysis of media, and social context analysis. Event recognition refers to the identification of simple and composite events that satisfy some pattern (e.g. the identifications of an emergency situation, of a protest or demonstration). Machine learning techniques are commonly used for this task, as well as methods for reducing the uncertainty in the identification process. Location analysis leverages on geotagged media as well as volunteered geographic information and can help assess the proximity of information to events and credibility of information (consider, for example, the case of a newscaster reporting immediately from an event location). Social interactions around content (e.g., comments) carry a rich information context, which, if appropriately analysed, can yield deep insights into the impact that media content has on users, and reveal popularity, and reputation modalities.