Report from Dagstuhl Seminar 17301

User-Generated Content in Social Media

Edited by

Tat-Seng Chua[1], Norbert Fuhr[2], Gregory Grefenstette[3], Kalervo Järvelin[4], and Jaakko Peltonen[5]

1 **National University of Singapore, SG,** chuats@comp.nus.edu.sg
2 **Universität Duisburg-Essen, DE,** norbert.fuhr@uni-due.de
3 **IHMC – Paris, FR,** ggrefenstette@ihmc.us
4 **University of Tampere, FI,** kalervo.jarvelin@uta.fi
5 **Aalto University & University of Tampere, FI,** jaakko.peltonen@uta.fi

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17301 "User-Generated Content in Social Media". Social media have a profound impact on individuals, businesses, and society. As users post vast amounts of text and multimedia content every minute, the analysis of this user generated content (UGC) can offer insights to individual and societal concerns and could be beneficial to a wide range of applications. In this seminar, we brought together researchers from different subfields of computer science, such as information retrieval, multimedia, natural language processing, machine learning and social media analytics. We discussed the specific properties of UGC, the general research tasks currently operating on this type of content, identifying their limitations, and imagining new types of applications. We formed two working groups, WG1 "Fake News and Credibility", WG2 "Summarizing and Story Telling from UGC". WG1 invented an "Information Nutrition Label" that characterizes a document by different features such as e.g. emotion, opinion, controversy, and topicality; For computing these feature values, available methods and open research issues were identified. WG2 developed a framework for summarizing heterogeneous, multilingual and multimodal data, discussed key challenges and applications of this framework.

Seminar July 23–28, 2017 – http://www.dagstuhl.de/17301

1998 ACM Subject Classification H Information Systems, H.5 Information Interfaces and Presentation, H.5.1 Multimedia Information Systems, H.3 Information Storage and Retrieval, H.1 Models and principles, I Computing methodologies, I.2 Artificial Intelligence, I.2.6 Learning, I.2.7 Natural language processing, J Computer Applications, J.4 Social and behavioural sciences, K Computing Milieux, K.4 Computers and Society, K.4.1 Public policy issues

Keywords and phrases social media, user-generated content, social multimedia, summarisation, storytelling, fake-news, credibility, AI

Digital Object Identifier 10.4230/DagRep.7.7.110

Edited in cooperation with Nicolás Díaz Ferreyra

1 Executive Summary

Norbert Fuhr
Tat-Seng Chua
Gregory Grefenstette
Kalervo Järvelin
Jaakko Peltonen

Social media play a central role in many people's lives, and they also have a profound impact on businesses and society. Users post vast amounts of content (text, photos, audio, video) every minute. This user generated content (UGC) has become increasingly multimedia in nature. It documents users' lives, revealing in real time their interests and concerns and activities in society. The analysis of UGC can offer insights to individual and societal concerns and could be beneficial to a wide range of applications, for example, tracking mobility in cities, identifying citizen's issues, opinion mining, and much more. In contrast to classical media, social media thrive by allowing anyone to publish content with few constraints and no oversight. Social media posts thus show great variation in length, content, quality, language, speech and other aspects. This heterogeneity poses new challenges for standard content access and analysis methods. On the other hand, UGC is often related to other public information (e.g. product reviews or discussion of news articles), and there often is rich contextual information linking, which allows for new types of analyses.

In this seminar, we aimed at discussing the specific properties of UGC, the general research tasks currently operating on this type of content, identifying their limitations and lacunae, and imagining new types of applications made possible by the availability of vast amounts of UGC. This type of content has specific properties such as presentation quality and style, bias and subjectivity of content, credibility of sources, contradictory statements, and heterogeneity of language and media. Current applications exploiting UGC include sentiment analysis, noise removal, indexing and retrieving UGC, recommendation and selection methods, summarization methods, credibility and reliability estimation, topic detection and tracking, topic development analysis and prediction, community detection, modeling of content and user interest trends, collaborative content creation, cross media and cross lingual analysis, multi-source and multi-task analysis, social media sites, live and real-time analysis of streaming data, and machine learning for big data analytics of UGC. These applications and methods involve contributions from several data analysis and machine learning research directions.

This seminar brought together researchers from different subfields of computer science, such as information retrieval, multimedia, natural language processing, machine learning and social media analytics. After participants gave presentations of their current research orientations concerning UGC, we decided to split into two Working Groups: (WG-1) Fake News and Credibility, and (WG-2) Summarizing and Storytelling from UGC.

WG-1: Fake News and Credibility

WG-1 began discussing the concept of Fake News, and we arrived at the conclusion that it was a topic with much nuance, and that a hard and fast definition of what was fake and what was real news would be hard to define. We then concentrated on deciding what elements of Fake (or Real) News could be calculated or quantified by computer. This led us to construct a list of text quality measures that have or are being studied in the Natural Language Processing community: Factuality, Reading Level, Virality, Emotion, Opinion, Controversy, Authority, Technicality, and Topicality. During this discussion, WG-1 invented and mocked up what we called an Information Nutrition Label, modeled after nutritional labels found on most food products nowadays. We feel that it would be possible to produce some indication of the "objective" value of a text using the above nine measures. The user could use these measures to judge for themselves whether a given text was "fake" or "real". For example, a text highly charged in Emotion, Opinion, Controversy, and Topicality might be Fake News for a given reader. Just like with a food nutritional label, a reader might use the Information Nutritional Label to judge whether a given news story was "healthy" or not.

WG-1 split into further subgroups to explore whether current status of research in the nine areas: Factuality, Reading Level, etc. For each topic, the subgroups sketched out the NLP task involved, found current packages, testbeds and datasets for the task, and provided recent bibliography for the topic. Re-uniting in one larger group, each subgroup reported on their findings, and we discussed next steps, envisaging the following options: a patent covering the idea, creating a startup that would implement all nine measures and produce a time-sensitive Information Nutritional Label for any text submitted to it, a hackathon that would ask programmers to create packages for any or all of the measurements, a further workshop around the Information Nutrition label, integration of the INL into teaching of Journalists, producing a joint article describing the idea. We opted for the final idea, and we produced a submission (also attached to this report) for the Winter issue of the SIGIR (Special Interest Group on Information Retrieval) Forum[1].

WG-2: Summarizing and Storytelling from UGC

WG-2 set out to re-examine the topic of summarization. Although this is an old topic, but in the era of user-generated content with accelerated rates of information creation and dissemination, there is a strong need to re-examine this topic from the new perspectives of timeliness, huge volume, multiple sources and multimodality. The temporal nature of this problem also brings it to the realm of storytelling, which is done separately from that of summarization. We thus need to move away from the traditional single source document-based summarization, by integrating summarization and storytelling, and refocusing the problem space to meet the new challenges.

We first split the group into two sub-groups, to discuss separately: (a) the motivations and scopes, and (b) the framework of summarization. The first sub-group discussed the sources of information for summarization including, the user-generated content, various authoritative information sources such as the news and Wikipedia, the sensor data, open data and proprietary data. The data is multilingual and multimodal, and often in real time. The group then discussed storytelling as a form of dynamic summarization. The second group examined the framework for summarization. It identified the key pipeline processes comprising of: data ingestion, extraction, reification, knowledge representation, followed by story generation. In particular, the group discussed the roles of time and location in data, knowledge and story representation.

[1] http://sigir.org/forum/

Finally, the group identified key challenges and applications of the summarization framework. The key challenges include multi-source data fusion, multilinguality and multimodality, the handling of time/ temporality/ history, data quality assessment and explainability, knowledge update and renewal, as well as focused story/ summary generation. The applications that can be used to focus the research includes event detection, business intelligence, entertainments and wellness. The discussions have been summarized into a paper entitled "Rethinking Summarization and Storytelling for Modern Social Multimedia". The paper is attached along with this report. It has been submitted to a conference for publication.

2 Contents

3 Overview of Talks

3.1 NLP Approaches for Fact Checking and Fake News Detection

Andreas Hanselowski (TU Darmstadt, DE)

In the past couple of years, there has been a significant increase of untrustworthy information on the web, such as fake news articles. In order to validate the false information circulating on the web, fact-checking became an essential tool, and today, there are numerous websites such as fullfact.org, politifact.com, and snopes.com devoted to the problem. Although manual fact-checking is an important instrument in the fight against false information, it cannot solve the problem entirely. The large number of fake news articles being generated at a high rate cannot be easy detected in debunked by human fact checkers. Many of the upcoming issues of manual claim validation can be addressed by automated fact-checking, as it would allow a large number of articles to be validated in real time as they appear on the web. The problem of tackling false information on the web can be divided into two different problem settings, that is, into fake news detection and automated fact-checking.

In the seminar, machine learning approaches for both problem-setting have been presented. For fake news detection, the problem of stance classification was discussed, as it was posed in the Fake News Challenge. For the solution of the problem a multilayer perception, which is based on linguistic features, was introduced. For the automated fact-checking, a comprehensive framework was presented, in which the problem is divided into a number of steps. In the first step, web documents are retrieved, which contain information required for the resolution of a given claim. Next, evidence in the web documents are identified, which explicitly support or contradict the claim. In the last step, the actual claim validation is performed, whereby the identified evidence and web documents serve as a basis.

3.2 User-Generated Content and Privacy Risks: From Regrets to Preventative Technologies

Nicolas Diaz-Ferreyra (Universität Duisburg-Essen, DE)

URL https://doi.org/10.1007/978-3-319-66808-6_7

User-generated content very often enclose private and sensitive information. When such information reaches an unintended audience, it can derivate in unwanted incidents for the users like job loss, reputation damage, or sextortion along with a feeling of regret. Preventative technologies aim to help users to bypass the potential unwanted incidents of online-self disclosure by raising awareness on the content being shared. However, in order to engage with the users, such technologies should follow some basic design principles. First, preventative technologies should be adaptive on the users' privacy attitudes and intentions. Second, they should generate a visceral connection between users and their private data. Finally, such technologies should provide supportive guidance to the users informing about possible actions that can help them to protect their privacy. This talk aim to discuss the role of regrettable user-generated content in the development of adaptive, visceral and supportive preventative technologies. Particularly, how privacy heuristics can be extracted from regrettable experiences and integrated later on into the design of awareness mechanisms.

3.3 Social Media Retrieval with Contradictions and Credibility

Norbert Fuhr (Universität Duisburg-Essen, DE)

User comments in Social Media – e.g. reviews of products or services – often contradict each other, and they also may vary in terms of credibility. In order to aggregate these comments for the purpose of retrieval, we propose to apply a number of logic-based concepts: 1) While today's retrieval methods mostly use an implicit closed world assumption, an open world assumption allows to distinguish between missing information and explicit negation. 2) For handing contradictions, a four-valued logic also contains the truth values 'inconsistent' and 'unknown'. 3) A possible worlds semantics models an agent's belief over possibly contradictory statements as a probability distribution over different worlds, where this distribution can be used for representing credibility of statements. While these logic-based formalisms are well developed, a major challenge in their application is the development of appropriate indexing methods for creating the representations needed for the application of the logic-based models.

3.4 Multilingual Aspects of User-Generated Content

Tatjana Gornostaja (tilde – Riga, LV)

My name is Tatjana Gornostaja and I present on behalf of the company Tilde I have been working for more than 10 years with my background in terminology (knowledge management) and translation (human and automated). Tilde is specialising in natural language data (text and speech) processing with the focus on small languages with scarce resources and rich grammar, developing innovative products and services of machine translation, terminology management, speech analysis, virtual assistants and operating in the three Baltic countries – Estonia, Latvia, Lithuania with the headquarters in Riga.

A huge amount of content is generated by users on the Internet (YouTube, Facebook, Instagram, Twitter, WordPress etc.) in different languages. According to statistics, more than 40% of Europeans speak only their native language and more than 60% do not speak English well enough to consume the content published on the Internet, which is predominantly in English. However, if you talk to a man in a language he understands – that goes to his head, if you speak to him in his own language – that goes to his heart (Nelson Mandela). This is our motto for the products and services we provide to our users (public administration, business, academia – locally and globally) to help them to communicate successfully.

The latest advancements have been integrated into popular platforms recently:

- best AI-powered machine translation for Latvian on Twitter
- virtual assistants (multiplication table for children, currency exchange and travel guides for adults) on Facebook Messenger and Skype

- speech and text processing systems for government services, including the support to the European Presidency in Latvia (past), Estonia (ongoing), Bulgaria and Austria (upcoming)
- speech processing mobile applications for Latvian for people with vision impairment and dyslexia.

Being proud and honoured for the invitation and participation in the Schloss Dagstuhl seminar, I am grateful to its organisers for this excellent event with outstanding presentations and inspiring discussions. With our competences, expertise and experience in more than 25 international research, development and innovation projects, with a wide range of more than 60 partners worldwide we are open for collaboration and new ideas to connect people speaking different languages across borders.

3.5 Spam in User-Generated Content

Gregory Grefenstette (IHMC – Paris, FR)

User Generated Content (UGC) provides rich data for understanding language use, user opinion, and many other uses. But researchers should be aware of the wide variety of spam that appears in this data. False blogs can generate well structured but random text to hide pointers to money-making sites. Comments may contain generic messages also to create links to spam pages. It is easy to create false users, false reviews, false likes for any social network. We examine some of these problems and show some ways to detect spam and false users, so that researchers know that this noise exists in their UGC.

3.6 Cross-Modal Recommendation: From Shallow Learning to Deep Learning

Xiangnan He (National University of Singapore, SG)

Recommender systems play a central role in the current user-centered Web. Many customer-oriented online services rely on recommender systems or advertising systems to earn money. The de-facto technique to build a personalized recommender system is collaborative filtering, which predicts a user's preference based on the historical user-item interactions. Besides the interaction data, there are also rich side information available, such as user demographics (e.g., gender, age), item attributes (e.g., textual description and visual images), and various contexts. The key research here is how to effectively leverage all relevant information available to build a better recommender system.

In this talk, I first present the existing and widely used techniques for building a generic recommender that model various information, including Logistic Regression (plus GBDTs), Factorization Machines (FMs), and Tensor Decomposition. Then, I briefly introduce recent deep learning solutions for recommendation, such as the Google's Wide&Deep and Microsoft's Deep Crossing. Lastly, I introduce our recently proposed neural recommendation solutions, Neural FMs and Attentional FMs.

3.7 Altmetrics and Tweeting Behavior of Scientists

Isabella Peters (ZBW – Dt. Zentralbib. Wirtschaftswissenschaften, DE)

The talk introduced altmetrics, the field of research evaluation by means of social media signals for scholarly products. Examples for ongoing research were given – for example the relationships between citations, the journal impact factor and altmetrics were discussed. It was shown that altmetrics are influenced by characteristics of the scholarly product, e.g., its popularity, the language in which it was published, and whether it was authored by known authors. Moreover, altmetrics are strongly dependent on the context the social media was used. The talk concluded that responsible use of altmetrics calls for better understanding of how, why, when and by whom social media signals are produced in order to draw correct conclusions from pure numbers.

Further reading at DBLP: conf/isiwi/NurediniP15 journals/it/HausteinLTAP14

3.8 People Analytics with User-Generated Content

Rianne Kaptein (Crunchr – Amsterdam, NL)

The relatively new field of people analytics has become a hot topic in organizations of all sizes. People Analytics, also known as HR analytics, refers to the method of analytics that can help managers and executives make decisions about their employees or workforce. Organizations are reaching out to learn more about predictive analytics in order to improve organizational effectiveness. Significant resources are devoted to identify talented employees and to develop them into future leaders. However, current people management processes are subjective and mostly retrospective. For example, emerged talent is identified in hindsight and valuable years of development are missed out on. Talented employees perceive this as being undervalued and might eventually leave the company. Other people analytics tasks include: Workforce planning, Succession planning, Recruitment optimization, Team composition, Predicting employee turnover and Employee engagement analysis. The main sources for user generated content on HR data are LinkedIn and Glassdoor. These sites include information on: resumes, work experience, skills, connections, company reviews, etc. Challenges when using this content are privacy, identity resolution and data sparseness. To overcome these challenges we do not identify individuals, but only analyze on the company level or anonymized data.

User generated content on Social Media often consist of a large number of short texts (e.g. survey question responses, tweets, Facebook Posts, forum posts). HR staff has problems with information overload – often there are too many messages to read. So we aim to give an overview or summary of the relevant/interesting responses. These summaries can also be coupled with sentiment analysis.

3.9 Detecting Malicious Activities in Community Question Answering Platforms

Yiqun Liu (Tsinghua University – Beijing, CN)

With Community Question Answering (CQA) evolving into a quite popular method for information seeking and providing, it also becomes a target for spammers to disseminate promotion campaigns. Although there are a number of quality estimation efforts on the CQA platform, most of these works focus on identifying and reducing low-quality answers, which are mostly generated by impatient or inexperienced answerers. However, a large number of promotion answers appear to provide high-quality information to cheat CQA users in future interactions. Therefore, most existing quality estimation works in CQA may fail to detect these specially designed answers or question-answer pairs. In contrast to these works, we proposed two methods for detecting spamming activities on CQA platforms. For individual spamming activity detection, we focus on the promotion channels of spammers, which include (shortened) URLs, telephone numbers and social media accounts. Spammers rely on these channels to connect to users to achieve promotion goals so they are irreplaceable for spamming activities. We propose a propagation algorithm to diffuse promotion intents on an "answerer-channel" bipartite graph and detect possible spamming activities. A supervised learning framework is also proposed to identify whether a QA pair is spam based on propagated promotion intents. Experimental results based on more than 6 million entries from a popular Chinese CQA portal show that our approach outperforms a number of existing quality estimation methods for detecting promotion campaigns on both the answer level and QA pair level. For collusive spamming activity detection, we propose a unified framework to tackle the challenge. First, we interpret the questions and answers in CQA as two independent networks. Second, we detect collusive question groups and answer groups from these two networks respectively by measuring the similarity of the contents posted within a short duration. Third, using attributes (individual-level and group-level) and correlations (user-based and content-based), we proposed a combined factor graph model to detect deceptive Q&As simultaneously by combining two independent factor graphs. With a large-scale practical data set, we find that the proposed framework can detect deceptive contents at early stage, and outperforms a number of competitive baselines.

3.10 Social Media and e-Commerce

Marie-Francine Moens (KU Leuven, BE)

The lecture has focused on representation learning of social media content and was illustrated with two use cases: Bridging the language of consumers and product vendors and bridging language and vision for cross-modal fashion search.

In the first use case, we have focused on linking content (textual descriptions of pins in Pinterest to webshops). We have explained the problem of linking information between different usages of the same language, e.g., colloquial and formal "idioms" or the language of consumers versus the language of sellers. For bridging these languages, we have trained a multi-idiomatic latent Dirichlet allocation model (MiLDA) on product descriptions aligned with their reviews.

In the second use case, we have proposed two architectures to link visual with textual content. The first architecture uses a bimodal latent Dirichlet allocation topic model to bridge between these two modalities. As a second architecture, we have developed a neural network which learns intermodal representations for fashion attributes. Both resulting models learn from organic e-commerce data, which is characterised by clean image material, but noisy and incomplete product descriptions. We have demonstrated two tasks: 1) Given a query image (without any accompanying text), we retrieve textual descriptions that correspond to the visual attributes in the visual query; and 2) given a textual query that expresses an interest in specific visual characteristics, we retrieve relevant images (without leveraging textual metadata) that exhibit the required visual attributes. The first task is especially useful to manage product image collections by online stores who might want to automatically organise and mine predominantly visual items according to their attributes without human input. The second task allows users to find product items with specific visual characteristics, in the case where there is no text available describing the target image.

3.11 Learning from Social Media and Contextualisation

Josiane Mothe (University of Toulouse, FR)

Main reference Idriss Abdou Malam, Mohamed Arziki, Mohammed Nezar Bellazrak, Farah Benamara, Assafa El Kaidi, Bouchra Es-Saghir, Zhaolong He, Mouad Housni, Véronique Moriceau, Josiane Mothe, Faneva Ramiandrisoa: "IRIT at e-Risk. CLEF", Working Notes of CLEF 2017, CEUR-WS.org, 2017.
URL http://ceur-ws.org/Vol-1866/paper_135.pdf

Social media can be a rich source of information either to extract some trends (models) or peculiarities (weak signals). We focused in this talk on early depression detection from social media posts using machine learning techniques and presented some results. We also proposed to use the same type of model to detect and extract locations from short posts when user localisation is not available. Finally, we mentioned our current work on tweet contextualisation that aims helping users to understand short texts.

3.12 Machine Learning for Analysis of Hierarchical Conversation Forums

Jaakko Peltonen (Aalto University, FI)

In many domains, user generated textual data arise as hierarchically organised document sets. In particular, online discussion often occurs as conversation threads in online message forums and other social media platforms having a prominent hierarchical organisation, with multiple levels of sections and subsections. Modeling the online discussions is important for studies of discussion behavior, for tracking trends of consumer interests, and analysis of brands and advertising impact.

Machine learning methods can help to analyse latent topics within such content, which can then be used for predictive and exploratory tasks, such as analysis of trends, analysis of user interest and sentiment across different types of content, recommendation of discussion content or targeting of advertising, and for intelligent interfaces to browse and participate in discussions. The hierarchical structure of online forums is designed to cover a subset of prototypical user interests, and could help build better models of content; however, most available machine learning methods cannot fully take the hierarchy into account in modelling.

In our recent work we have developed methods for taking this hierarchical structure into account in modelling latent topics of discussion, including probabilistic nonparametric topic models of the hierarchical content and interactive exploratory interfaces based on dimensionality reduction that reveal how the variety of discussion content is related to the hierarchy. We are further developing methods to model how individual users and populations of users visit content across the hierarchy.

3.13 Multimedia in Data Science: Bringing Multimedia Analytics to the Masses

Stevan Rudinac (University of Amsterdam, NL)

In this talk, using urban computing as a case study, we advocate that Multimedia community should embrace data science. Increased availability of open data in relation to various neighbourhood statistics such as demographics, transportation and services, made analysing and modelling processes in the city significantly easier. However, useful information about the problems a city is facing with may be also extracted from spontaneously captured social multimedia, participatory data and wearable technology. The examples from our work on interactive venue recommendation, discovering functional regions of the city, analysing liveability of the neighbourhoods and empowering local urban communities, demonstrate that multimedia analytics can be successfully deployed on such heterogeneous data for solving important societal problems. We further show that multimedia analytics and, in particular, interactive learning can be facilitated even on very large collections with 100 million user-generated images and associated annotations. Perhaps more importantly, it may be a possible solution for the imperfections in the automatic analysis techniques and facilitate easier technology adoption by keeping the user in control. Finally, we touch the topics of data reliability, privacy and ethics.

3.14 User- and Culture-Aware Models for Music Recommender Systems

Markus Schedl (Universität Linz, AT)

Nowadays, music aficionados generate millions of listening events every day and share them via services such as Last.fm or Twitter. In 2016, the LFM-1b dataset http://www.cp.jku.at/datasets/LFM-1b containing more than 1 billion listening events of about 120,000 Last.fm users has been released to the research community and interested public. Since then, we performed various data analysis and machine learning tasks on these large amounts of user and listening data. The gained insights helped to develop new listener models and integration them into music recommender systems, in an effort to increase personalization of the recommendations. In this talk, the focus is on the following research directions, which we are currently pursuing: (i) analyzing music taste around the world and distilling country clusters, (ii) quantifying listener and country mainstreaminess, (iii) music recommendation tailored to listener characteristics, and (iv) predicting country-specific genre preferences from cultural and socio-economic factors.

3.15 Changing our Mind: Correlations of Media in Online Collaboration Systems

David Ayman Shamma (CWI – Amsterdam, NL)

As humans, we create a lot of data and we change our minds. Sometimes we learn and grow; sometimes we were just wrong. In particular, we see these edits and changes in online user generated social systems. However, rarely are these changes accounted for when we index, recommend, and classify. In this talk, I illustrate, using historical Wikipedia associations, how community use and abuse changes the semantics and meaning of the images we use. Further, I assert we need to know why people make and change annotations as it changes how we build artificial intelligence systems for user-generated media.

3.16 Social Media: A Narcissic Form of Lifelogging?

Alan Smeaton (Dublin City University, IE)

Main reference Cathal Gurrin, Alan F. Smeaton, Aiden R. Doherty: "LifeLogging: Personal Big Data", Foundations and Trends in Information Retrieval, Vol. 8(1), pp. 1–125, 2014.
URL http://dx.doi.org/10.1561/1500000033

I present the state of work in lifelogging, first person digital ethnography, using off the shelf wearable sensors coupled with data from public sources. The talk focuses a lot on using wearable cameras and the various kinds of behaviour and activities can than be automatically extracted from such wearable camera images. I also present our lab's work on new kinds of wearable sensors for glucose levels or sweat composition. The final argument of the presentation is that narcissic social media posts can be replaced by something extracted from human lifelogs.

3.17 User-Generated Content: An Adversarial Perspective

Benno Stein (Bauhaus-Universität Weimar, DE)

In this talk I will present research questions and results of selected adversarial analytics where our research group has been working on in the recent past: Hyperpartisan news (in Blogs), Clickbait (in Twitter) [80], Wikipedia Vandalism [81], Offensive Language in comments. The talk will point out the different natures of the adversarial incidents, which in turn give raise for different countermeasures: clickbait can be addressed with automatically formulating and submitting suited questions queries whose results are presented along the clickbait text. Similarly, "alternative" facts or fake news can be countered by consulting an argument search engine [82]. Aside from theoretical backgrounds the talk will provide demonstrations of recently developed technology.

3.18 An Anatomy of Online Video Popularity

Lexing Xie (Australian National University – Canberra, AU)

How did a video go viral? Or will it go viral, and when? These are some of the most intriguing yet difficult questions in social media analysis. I will cover a few recent results from my group on understanding and predicting popularity, especially for YouTube videos. I will start by describing a unique longitudinal measurement study on video popularity history, and introduce popularity phases, a novel way to describe the evolution of popularity over time. I will then discuss a physics-inspired stochastic model that connects exogenous stimuli and endogenous responses to explain and forecast popularity. This, in turn, leads to a set of novel metrics for forecasting expected popularity gain per share, the time it takes for such effects to unfold, and sensitivity to promotions.

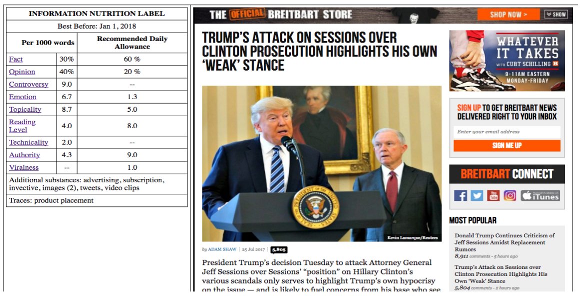

Figure 1 Mockup of the envisaged information nutrition label.

4 Working groups

4.1 An Information Nutritional Label for Online Documents

Norbert Fuhr (Universität Duisburg-Essen, DE), Anastasia Giachanou (University of Lugano, CH), Gregory Grefenstette (IHMC – Paris, FR), Iryna Gurevych (TU Darmstadt, DE), Andreas Hanselowski (TU Darmstadt, DE), Kalervo Järvelin (University of Tampere, FI), Rosie Jones (Microsoft New England R&D Center – Cambridge, US), Yiqun Liu (Tsinghua University – Beijing, CN), Josiane Mothe (University of Toulouse, FR), Wolfgang Nejdl (Leibniz Universität Hannover, DE), Isabella Peters (ZBW Dt. Zentralbib. Wirtschaftswissenschaften, DE), and Benno Stein (Bauhaus-Universität Weimar, DE)

With the proliferation of online information sources, it has become more and more difficult to judge the trustworthiness of news found on the Web. The beauty of the web is its openness, but this openness has lead to a proliferation of false and unreliable information, whose presentation makes it difficult to detect. It may be impossible to detect what is "real news" and what is "fake news" since this discussion ultimately leads to a deep philosophical discussion of what is true and what is false. However, recent advances in natural language processing allow us to analyze information objectively according to certain criteria (for example, the number of spelling errors). Here we propose creating an "information nutrition label" that can be automatically generated for any online text. Among others, the label provides information on the following computable criteria: factuality, virality, opinion, controversy, authority, technicality, and topicality.

4.1.1 Introduction

The 2016 American presidential elections were a source of growing public awareness of what has been termed "fake news". In a nutshell, the term is used to describe the observation that "in social media, a certain kind of 'news' spread much more successfully than others, and, that these 'news' stories are typically extremely one-sided (hyperpartisan), inflammatory, emotional, and often riddled with untruths" [71].

Claims in news can take various forms. In the form of a verifiable assertion ("The density of ice is larger than the density of water.") we have a fact checking situation, which can be clarified given access to online dictionaries or encyclopedias. In the form of a non-verifiable or not easily verifiable assertion ("Hillary Clinton is running a child sex ring out of a D.C.-area pizza restaurant.", "Marijuana is safer than alcohol or tobacco.") one has to take a stance, i.e., the reader has to decide whether she believes the claim or not. Such a decision can neither universally nor uniquely be answered by means of a knowledge base but is to be clarified on an individual basis and may undergo change over time.

To help the online information consumer, we propose an Information Nutrition Label, resembling nutrition fact labels on food packages. Such a label describes, along a range of agreed-upon dimensions, the contents of the product (an information object, in our case) in order to help the consumer (reader) in deciding about the consumption of the object. The observations above however imply technical, but in particular self-imposed ethical limitations of our envisaged concept:

> *(manifest) It is not our intention to say what is true or what is fault, right or wrong, and in particular not what is good or bad. That is, an Information Nutrition Label is not a substitute for a moral compass.*

Thus, as technical consequence, we neither propose a system that would state what is true or what is false, right or wrong, and in particular not what is good or bad. Ultimately, it is up to the consumer to consult the information nutrition label and to decide whether to consume the information or not. Aiming at aiding the consumer's decision making process, we see various technical uses as well as societal impacts of our Information Nutrition Label:
- personalized relevance ranking for search engine results
- information filtering according to personal preferences
- machine-based fake news detection
- learning and teaching of information assessment
- raising awareness and responsibility about deciding what to read.

4.1.2 An Information Nutrition Label

Of course, the assessment of information is not a new discipline—recall the large body of research related to the concept of "information quality", for which Levis et al. provide a useful overview [66]. While there is no unique definition for the concept, information quality is usually interpreted in terms of utility, namely as the "fitness for use in a practical application" [78]. Note that our paper will neither reinterpret nor extend this quality concept; instead, we are aiming at a practical means to ease information consumption and meta reasoning when given an online document by breaking down a quality judgment into smaller, measurable components.

We consider the Wikipedia quality endeavour as the most related precursor to our proposal. Aside from its rather informal quality guidelines, Wikipedia has formalized its quality ideal with the so-called featured article criteria[2], and, even more important, distinguishes more than 400 quality flaws to spot article deficits [53]. In particular, the machine-based analysis of Wikipedia articles to detect flaws in order to assess article quality [54] corresponds closely to our idea of computing the separate dimensions of an information nutrition label. However, because of our use case, the nutrition label dimensions as well as their computation differs from the Wikipedia setting.

The following subsubsections describe measurable qualities that may be included in such an information nutrition label and that we consider valuable in order to assess the

[2] Wikipedia, "Featured articles," last modified February 19, 2017, http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

nutrient content of an information object. Each of these categories have been the subject of experimentation in the natural language processing, information retrieval, or web sciences communities:

- factuality
- readability
- virality
- emotion
- opinion
- controversy
- authority / credibility / trust
- technicality
- topicality

In the next subsubsections we will define these aspects and describe the relationship they have with an information nutrition label Moreover, tasks and methods explain what practical steps may need to be taken into account to automatize the measurement of these qualities for insertion into the nutrition label. We consider the task descriptions as broad avenues from which further research may take off.

4.1.3 Factuality

4.1.3.1 Task for Factuality Assessment

The task of determining the level of commitment towards a predicate in a sentence according to a specific source, like the author, is typically addressed as factuality prediction ([73]). Lexical cues, such as modals *will*, *shall*, *can* indicate the confidence of the source whether a proposition is factual. However, in contrast to a binary decision, the underlying linguistic system forms a continuous spectrum ranging from factual to counterfactual ([73]). Thus, for the assessment of the factuality for the whole document, one needs to compute the average factuality of all the propositions contained in the text.

Since we are not planning to judge the truthfulness of the statements in a given text, as it is attempted in the domain of automated fact checking, we are only interested in determining whether a statement is factual from the perspective of the author. The issue of whether the statements in the documents are controversial and may therefore not be reliable, is discussed in subsubsection 4.1.8 about Controversy.

4.1.3.2 Methods for Factuality Assessment

For factuality prediction rule-based approaches as well as methods based on machine learning have been developed.

The De Facto factuality profiler [74] and the TruthTeller algorithm [68] are rule-based approaches, which assign discrete scores of factuality to propositions. In the process, dependency parse trees are analyzed top-down and the factuality score is altered whenever factuality affecting predicates or modality and negation cues are encountered.

A machine learning based approach has been applied to factuality prediction in [65]. The authors used a support vector machine regression model to predict continuous factuality values from shallow lexical and syntactic features such as lemmas, part-of-speech tags, and dependency paths.

The rule-based approach has been combined with the machine learning based method in [76]. Thereby, the outputs from TruthTeller were used as linguistically-informed features for a support vector machine regression model in order to predict the final factuality value.

4.1.3.3 Data sets for Factuality Assessment

There are a number of annotation frameworks, which have been suggested to capture the factuality of statements. On the basis of the suggested annotation schemes, a number of data sets have been constructed.

Fact-Bank [73] is a corpus which was annotated discretely by experts according to different classes of factuality: Factual, Probable, Possible, Unknown. In this corpus, factuality has been assessed with respect to the perspective of the author or discourse-internal sources.

The MEANTIME corpus was introduced in [69] and was also annotated discretely by expert annotators. The propositions have been classified as Fact / Counterfact, Possibility (uncertain), Possibility (future) with respect to the author's perspective.

The UW corpus [65] was annotated on the basis of a continuous scale ranging from -3 to 3. The annotation was performed by crowd workers who judged the factuality score from the author's perspective.

In [76], the annotation schemes of the three different corpora have been merged in order to combine the three data sets into one single large corpus. For this purpose, the discrete scales used for the Fact-Bank and MEANTIME corpora have been mapped to the continuous scale of the UW corpus.

4.1.3.4 Further reading for Factuality Assessment

1. Nissim Malvina, Paola Pietrandrea, Andrea Sanso, and Caterina Mauri. "Cross-linguistic annotation of modality: a data-driven hierarchical model." In Proceedings of the 9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, pp. 7-14. 2013.
2. O'Gorman Tim, Kristin Wright-Bettner, and Martha Palmer. "Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation." In Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016). 2016.
3. Ghia Elisa, Lennart Kloppenburg, Malvina Nissim, Paola Pietrandrea, and Valerio Cervoni. "A construction-centered approach to the annotation of modality." In Proceedings of the 12th ISO Workshop on Interoperable Semantic Annotation, pp. 67-74. 2016.
4. Guggilla Chinnappa, Tristan Miller, and Iryna Gurevych. "CNN-and LSTM-based Claim Classification in Online User Comments." In Proceedings of the COLING 2016.
5. Szarvas György, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. "Cross-genre and cross-domain detection of semantic uncertainty." Computational Linguistics 38, no. 2 (2012): 335-367.

4.1.4 Readability

4.1.4.1 Task for Readability Measurement

Readability is defined as "the ease with which a reader can understand a written text" [Wikipedia].

Readability can be measured by the accuracy of reading and the reading speed for the reader. Readability depends mainly on three categories of factors: writing quality, targeted audience and presentation.

Writing quality refers to the grammatical correctness of the text (morphology, syntax) such as taught in elementary schools [79]. Readability also depends on the *target audience* or in other words the level of educational background the reader needs to have to understand the text content (the complexity of its vocabulary and syntax, the rhetorical structure).

Finally, the *presentation* refers to typographic aspects like font size, line height, and line length [56] or visual aspects like color [64].

4.1.4.2 Methods for Readability Measurement

Collins-Thompson provides a recent state of the art summary of automatic text readability assessment [58]. Two main factors are used in readability measures: the familiarity of semantic units (vocabulary) and the complexity of syntax.

Automatic readability measures estimate the years of education or reading level required to read a given body of text using surface characteristics. The current measures are basically linear regressions based on the number of words, syllables, and sentences [70] [58].

Wikipedia presents a number of readability tests in their eponymous article that usually involve counting syllables, word length, sentence length, and number of words and sentences.[3]

Crossley et al. developed Coh-Metrix, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis [59]. De Clercq et al. proposed to use the crowd to predict text readibility [61].

Automatic methods have been developed for different languages as for Arabic [52], French [63], Polish [57], or Spanish [75] to cite a few.

4.1.4.3 Data sets for Readability Measurement

There are a number of data sets sets and sample demos for readability measurement. The data sets include:

- Text Exemplars and Sample Performance Tasks in Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects (183 pages). [Copyright and Permissions] Includes examples with different grades, genres (English only).[4]
- [62] mentions a collection of Weekly Reader extracts that may still be available.
- Math Webpage Corpus with Readability Judgments[5]

Sample demos:

- Readability[6] implemented by Andreas van Cranenburgh (*andreasvc* on github) calculates a number of standard reading level features, including Flesch, Kincaid and Smog (a descendent of an nltk_contrib package [7]). This package expects sentence-segmented and tokenized text. For English, van Cranenburgh recommends "tokenizer".[8] For Dutch, he recommends the tokenizer that is part of the Alpino parser[9]. There is also *ucto*[10], a general multilingual tokenizer. One can also use the tokenizer included in the Stanford NLP package.

3 https://en.wikipedia.org/wiki/Readability#Popular_readability_formulas, last access Oct. 11, 2017
4 http://www.corestandards.org/assets/Appendix_B.pdf
5 https://web.archive.org/web/*/http://wing.comp.nus.edu.sg/downloads/mwc
6 https://pypi.python.org/pypi/readability
7 https://github.com/nltk/nltk_contrib/tree/master/nltk_contrib/readability
8 http://moin.delph-in.net/WeSearch/DocumentParsing
9 http://www.let.rug.nl/vannoord/alp/Alpino/
10 http://ilk.uvt.nl/ucto

Test cases:

```
$ ucto -L en -n -s '' ''CONRAD, Joseph - Lord Jim.txt'' | readability
[...]
readability grades:
Kincaid:                4.95
ARI:                    5.78
Coleman-Liau:           6.87
FleschReadingEase:     86.18
GunningFogIndex:        9.4
LIX:                   30.97
SMOGIndex:              9.2
RIX:                    2.39
```

Other tools: Benchmark Assessor Live[11], and also see Further Reading, below.

4.1.4.4 Further reading for Readability Measuring

1. Flesch and Kincaid Readability tests, [12] and the Wikipedia article on Readability[13] (for several other readability formulas)
2. Heilman, Michael, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. "ombining lexical and grammatical features to improve readability measures for first and second language texts." In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 460-467. 2007.
3. Collins-Thompson, Kevyn. "Computational assessment of text readability: A survey of current and future research." ITL-International Journal of Applied Linguistics 165, no. 2 (2014): 97-135.
4. De La CHICA, Sebastian, Kevyn B. Collins-Thompson, Paul N. Bennett, David Alexander Sontag, and Ryen W. White. "Using reading levels in responding to requests." U.S. Patent 9,600,585, issued March 21, 2017.
5. Vajjala, Sowmya, and Detmar Meurers. "On the applicability of readability models to web texts." In Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pp. 59-68. 2013.
6. Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. "Frequent words improve readability and short words improve understandability for people with dyslexia." In IFIP Conference on Human-Computer Interaction, pp. 203-219. Springer, Berlin, Heidelberg, 2013. (excerpt: ... To determine how much individual queries differ in terms of the readability of the documents they retrieve, we also looked at the results for each query separately. Figure 4 shows the mean reading level of the Top-100 results for each of the 50 search queries...)
7. Newbold, Neil, Harry McLaughlin, and Lee Gillam. "Rank by readability: Document weighting for information retrieval." Advances in multidisciplinary retrieval (2010): 20-30. ("...Web pages can be, increasingly, badly written with unfamiliar words, poor use of syntax, ambiguous phrases and so on....")

[11] https://www.readnaturally.com/assessment-tools
[12] https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests
[13] https://en.wikipedia.org/wiki/Readability

8. Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. "A comparison of features for automatic readability assessment." In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 276-284. Association for Computational Linguistics, 2010.

4.1.5 Virality

In analyses of information objects and information flows on the internet the notion of "virality" is often stressed. This is especially true when discussed in the context of marketing and advertisement. Virality means that "information objects spread the way that viruses propagate. [Hence,v]irality has become a common way to describe how thoughts or information move through a human population, or the internet, social network sites in particular" [14]. The metaphor of the virus supports consideration of different properties that may influence the spread of information but that can also be used to quantify virality.

4.1.5.1 Task for Virality Detection

For the detection of virality in texts or other information objects four types of property sets have to be taken into account: a) the sender, b) the information object, c) the recipient, and d) the channel. The combination of these sets influences the speed with which a virus spreads and also determines how far it can reach. The major factors on the sender side are their popularity and authority, the size of their network, but also the amount of trust they receive from recipients. The recipient must be able to receive the information object and should not be immune to it, e.g. because they had the information object before. The information object itself is often admissable to many different types of recipients, for example, because of its short topical distance to knowledge the recipients already hold. The channel offers varying functionalities and allows for different ease of use to further spread the information object. Higher ease of use encourages the sharing of information objects, e.g., retweeting a tweet on Twitter. Moreover, the environment in which the information object spreads is of interest, too. It may have been influenced by a frame setting activity, i.e. bringing certain information to the awareness of many recipients, that increases the probability of recipients getting infected, e.g. because they search for this type of information. The virality of information objects could also be subject to within-platform as well as cross-platform properties.

4.1.5.2 Methods for Virality Detection

The determination of virality needs to operationalize all of these factors, especially with regard to the graph-like structure of the information flow. In social media, many signals can be used for this, e.g., number of likes, retweets, and comments, characteristics of followers, communities, or hashtags, or time of posting. Those factors build the ground for virality measurement. However, it is not only the quantity of these signals that may determine virality but also the speed with which information objects spread and how far they reach (e.g., when different communities are infected by the same information object).

[14] https://en.wikipedia.org/wiki/Viral_phenomenon

4.1.5.3 Tools and Data for Virality Detection

Examples for existing software that visualizes the spread of claims (i.e. Hoaxy) or that follows memes are provided by the Indiana University Network Science Institute (IUNI) and the Center for Complex Networks and Systems Research (CNetS)[15].

There are also several data sets available that can be used for training, for example viral images[16] or tweets[17], see also Weng *et al.* in the Further Reading subsubsection.

4.1.5.4 Further reading for Virality

1. Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. "Virality prediction and community structure in social networks." Scientific reports 3 (2013): 2522.
2. Weng, Lilian, and Filippo Menczer. "Topicality and impact in social media: diverse messages, focused messengers" PloS one 10, no. 2 (2015): e0118410.
3. Guerini, Marco, Carlo Strapparava, and Gözde Özbal. "Exploring Text Virality in Social Networks." In ICWSM. 2011.
4. Guille, Adrien and Hacid, Hakim and Favre, Cecile and Zighed, Djamel A. "Information diffusion in online social networks: A survey." ACM Sigmod Record 42.2 (2013): 17-28.

4.1.6 Emotion

4.1.6.1 Task for Emotion Detection

One characteristic of Fake News is that it may make an inflammatory emotional appeal to the reader. Emotional arguments often employ words that are charged with positive or negative connotations (such as *bold* or *cowardly*). Such language also appears in product and movie reviews.

The task here is to detect the sentences which are emotive in a document, and to calculate the intensity, the polarity and the classes of the affect words found there. The emotional impact of a document can either be averaged over the number of words, or be calculated by using some maximum value encountered [55].

4.1.6.2 Methods for Emotion Detection

As a sample method, an emotion detection method can include the following steps:
1. Divide document into sentences
2. Extract words, terms, negations, intensifiers, emoticons, parts of speech, punctuation from the sentence
3. Use these extracted items as features to classify the sentence
4. Identify which sentences carry emotion, and what emotion
5. Combine measures from all sentences to create a single emotion rating of the document.

4.1.6.3 Data sets for Emotion Detection

Data resources for emotion detection include sentiment lexicons and test/training data sets. Some of the former are:

---

[15] http://truthy.indiana.edu
[16] https://github.com/ArturoDeza/virality
[17] http://carl.cs.indiana.edu/data/#virality2013

- A list of Affect Lexicons[18] maintained by Saif Mohammad
- SenticNet[19]
- AFINN[20]
- List of affect resources[21] maintained by Bing Liu
- Affective Norms for English Words (ANEW) is a set of normative emotional ratings for 2,476 English words. We use the "valence" rating considering positive (respectively, negative) the ratings above (respectively, below) the mean.
- General Inquirer is a list of 1,915 words classified as positive, and 2,291 words classified as negative.
- MicroWNOp is a list of 1,105 WordNet synsets (cognitive synonyms) classified as positive, negative, or neutral.
- SentiWordNet assigns to each synset of WordNet (around 117,000) a positive and negative score determined by a diffusion process.
- Bias Lexicon is a list of 654 bias-related lemmas extracted from the edit history of Wikipedia [72]. Sentiment words are used as contributing features in the construction of this bias lexicon.

Test and training data sets include: Reviews;[22] Twitter in 15 languages;[23] Twitter and emotions;[24] Twitter tweets;[25] Blog sentences;[26] Facebook statuses, CNN, the New York Times, Guardian, BBC news, ABC news;[27] three emotional dimensions (Valence, Arousal and Dominance)[28]

4.1.6.4 Further reading for Emotion Detection

1. Valitutti, Alessandro, and Carlo Strapparava. "Interfacing WordNet-affect with OCC model of emotions."In The Workshop Programme, p. 16. 2010.[29]
2. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4 (2014): 1093-1113.
3. Giachanou, Anastasia, and Fabio Crestani. "Like it or not: A survey of twitter sentiment analysis methods." ACM Computing Surveys (CSUR) 49, no. 2 (2016): 28.
4. Cambria, Erik. "Affective computing and sentiment analysis." IEEE Intelligent Systems 31, no. 2 (2016): 102-107.
5. Tripathi, Vaibhav, Aditya Joshi, and Pushpak Bhattacharyya. "Emotion Analysis from Text: A Survey."[30]

[18] http://saifmohammad.com/WebPages/lexicons.html
[19] http://sentic.net/downloads/
[20] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
[21] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html
[22] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#datasets
[23] https://www.clarin.si/repository/xmlui/handle/11356/1054
[24] http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html
[25] http://www.sananalytics.com/lab/twitter-sentiment/
[26] https://inclass.kaggle.com/c/si650winter11/data
[27] https://github.com/minimaxir/interactive-facebook-reactions/tree/master/data
[28] https://github.com/JULIELab/EmoBank/tree/master/corpus
[29] https://source.opennews.org/articles/analysis-emotional-language/
[30] http://www.cfilt.iitb.ac.in/resources/surveys/emotion-analysis-survey-2016-vaibhav.pdf

Table 1 Examples of Fact vs Opinion sentences as taught to US Elementary School Children[a], along with a score which could be computed from them.

Sentence	Label
The first amendment includes the most misused freedom in our country, which is the freedom of the press.	`Opinionated`
The 18th amendment to the constitution prohibited the manufacture, sale, or transportation of alcohol.	`Fact`
The 16th amendment gave congress to collect taxes from American citizens, and they have been collecting way too many taxes ever since	`Opinionated`
Result	Opinion-Ratio = 2/3

[a] http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postest.html

4.1.7 Opinion

Opinion is an element of the text which reflects the author's opinion, and readers' opinions may differ. The output is a percentage, based on the fraction of words or sentences which are opinion, in contrast to facts. Authors of opinionated text may be surreptitiously pushing a certain viewpoint which is not explicitly expressed in the text.

4.1.7.1 Task for Opinion Detection

For the Information Nutrition Label, our task is to detect sentences that are opinionated, and calculate the percentage of opinionated sentences for entire text. Table 1 gives some examples of opinionated and factual sentences.

4.1.7.2 Existing Methods for Opinion Detection

There is software available for opinion detection. Here are some:
- OpeNER[31] "aims to be able to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on the texts[32]..."
- Opinion Finder[33], see Wilson et al, in Further Readings below.
- Opinion Sentence Finder[34]. See also Rajkumar et al., below.
- NLTK opinion lexicon reader[35].

4.1.7.3 Data sets for Opinion Detection

There are also data sets for opinion detection:
- Fact vs. opinion as taught to US Elementary School Children.[36] These examples have answers[37], too. The overall output score is the percent of sentences which contain opinions.

[31] http://www.opener-project.eu/
[32] http://www.opener-project.eu/getting-started/#opinion-detection
[33] http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/
[34] http://cse.iitkgp.ac.in/resgrp/cnerg/temp2/final.php
[35] http://www.nltk.org/_modules/nltk/corpus/reader/opinion_lexicon.html
[36] http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postest.html
[37] http://www.shsu.edu/txcae/Powerpoints/prepostest/fact1postans.html

- Bitterlemon collection 594 editorials about the Israel-Palestine conflict, 312 articles from Israeli authors and 282 articles from Palestinian authors.
- Opinion lexicon[38]
- Multi perspective question answering lexicon[39] corpus contains news articles and other text documents manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.).
- Arguing Lexicon[40]: includes patterns that represent arguing.

4.1.7.4 Further reading for Opinion Detection

1. Fact vs opinion as taught to US Elementary School Children[41]
2. Paul, Michael J., ChengXiang Zhai, and Roxana Girju. "Summarizing contrastive viewpoints in opinionated text." In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 66-76. Association for Computational Linguistics, 2010.
3. Yu, Hong, and Vasileios Hatzivassiloglou. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences." In Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 129-136. Association for Computational Linguistics, 2003. "classify sentences as fact / opinion using word n-grams, word polarity"
4. Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." In Proceedings of the 14th international conference on World Wide Web, pp. 342-351. ACM, 2005.
5. Wilson, Theresa, David R. Pierce, and Janyce Wiebe. "Identifying opinionated sentences." In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4, pp. 33-34. Association for Computational Linguistics, 2003.
6. Rajkumar, Pujari, Swara Desai, Niloy Ganguly, and Pawan Goyal. "A Novel Two-stage Framework for Extracting Opinionated Sentences from News Articles." In TextGraphs@ EMNLP, pp. 25-33. 2014.

4.1.8 Controversy

Controversy is a state of prolonged public dispute or debate, usually concerning a matter of conflicting opinion or point of view. The word was coined from the Latin controversia, as a composite of controversus – "turned in an opposite direction," from contra – "against" – and vertere – to turn, or versus (see verse), hence, "to turn against." The most applicable or well known controversial subjects, topics or areas are politics, religion, philosophy, parenting and sex (see Wikipedia articles in Further Reading, as well as Aharoni et al.) History is similarly controversial. Other prominent areas of controversy are economics, science, finances, culture, education, the military, society, celebrities, organisation, the media, age, gender, and race. Controversy in matters of theology has traditionally been particularly heated, giving rise to the phrase *odium theologicum*. Controversial issues are held as potentially divisive in a

[38] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#lexicon
[39] mpqa.cs.pitt.edu/corpora/mpqa_corpus/
[40] http://mpqa.cs.pitt.edu/lexicons/arg_lexicon
[41] http://teaching.monster.com/training/articles/2589-k-5-fact-versus-opinion

given society, because they can lead to tension and ill will, and as a result they are often considered taboo to be discussed in the light of company in many cultures.

Wikipedia lists some 2000 controversial issues.

4.1.8.1 Task for Controversy Detection

In its simplest form, for the Information Nutrition Label, we can calculate the number of controversial subjects in the text. A more evolved form would to calculate the density of controversial subjects in the text.

4.1.8.2 Methods for Controversy Detection

One method we can suggest for calculating the controversy of a text would be to look at those papers that implement Wikipedia featured article detection: they have to address the controversy flaw (the developed technology has parts that apply to non-Wikipedia articles as well). For topics that are covered by Wikipedia, determine the portion of reverts (after article editing), the so-called "edit wars" in Wikipedia. See the coverage measure (essay articles) below. Compute a number of features that hint controversy: topicality, retweet number and probability, query logs.

4.1.8.3 Data sets for Controversy Detection

Data Sources: Aharoni *at al.* (see further reading) describes a novel and unique argumentative structure dataset. This corpus consists of data extracted from hundreds of Wikipedia articles using a meticulously monitored manual annotation process. The result is 2,683 argument elements, collected in the context of 33 controversial topics, organized under a simple claim-evidence structure. The obtained data are publicly available for academic research.

The paper by Dori-Hacohen and Allan below also has a data set.

Test cases

Balance the number of pro and con arguments, using an argument search engine[42]. For queries/documents, which contain one of the controversial topics listed on the Wikipedia page, search/find documents that discuss (essay-like style) a topic. Choose documents appropriate for a specific reading level/background. Extract keywords/concepts and measure the overlap with controversial topics list (Wikipedia), debate portals, and the like.

4.1.8.4 Further reading for Controversy Detection

1. Wikipedia "Controversy" article [43]
2. Wikipedia list of controversial issues [44]
3. Examples of discussions of controversial topics can be found in the Scientific American[45] and on Plato[46].
4. Aharoni, Ehud, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. "A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics." In ArgMining@ACL, pp. 64-68. 2014.

[42] ttp://141.54.172.105:8081/
[43] https://en.wikipedia.org/wiki/Controversy
[44] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues
[45] https://www.scientificamerican.com/article/fact-or-fiction-vaccines-are-dangerous/
[46] https://plato.stanford.edu/entries/creationism/

5. Dori-Hacohen, Shiri, and James Allan. "Detecting controversy on the web." In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp. 1845-1848. ACM, 2013. "... Our approach maps a webpage to a set of Wikipedia articles, and uses the controversiality of those ... used two stop sets, the 418 INQUERY stop set [4] or a short, 35 term set ("Full" vs. ... 3. Handling non labeled data: We use two alterna- tives to "fill in the blanks" when labeled data ..."

4.1.9 Authority / Credibility / Trust

For the Information Nutrition Label, we consider trust and authority as synonyms that refer to a property of the source of a message, while credibility is an attribute of the message itself. On the Web, trust is assigned to a web site, while the different pages of the web site may be different in terms of credibility. When looking at a single document, users are most interested in its credibility; on the other hand, even experienced users judge credibility mainly based on their trust of the source. In the same way, for a system, however, it is easier to estimate the authority of a source (based on the information available), while there might be little document-specific evidence concerning its credibility.

4.1.9.1 Task for Authority

The task is to determine the authority or trust of the source of a document. Here we focus on Web sites and social media as sources.

4.1.9.2 Methods for Authority

For Web sites, a large number of methods for estimating authority have been proposed, of which we mention just a few:

- PageRank (Further Reading 1) is the most popular method for computing the importance of a Web site.
- Kleinberg's HITS algorithm (Further Reading 2) distinguishes between hub and authority scores.
- BrowseRank (Further Reading 3) computes the importance of a Web site by analysing user behavior data.
- Alexa Rank[47] measures Web site's popularity based solely on traffic to that site, in the form of a combined measure of unique visitors and page views of a website.

Recently, there also have been some approaches addressing the credibility of social media messages:

- Tweetcreed [4] is a Chrome browser extension computing a credibility score for a tweet using six types of features: meta-data, content-based simple lexical features, content-based linguistic features, author, external link URL's reputation, and author network.
- Sharriff et al. (Further Reading 5) aimed at estimating credibility perception of Twitter news considering features such as reader demographics, news attributes and tweet features.
- Popat et al. (Further Reading 6) presents a method for automatically assessing the credibility of claims in a message, which retrieves corresponding articles and models their properties such as the stance language style, their reliability, time information as well as their interrelationships.

[47] https://www.alexa.com

4.1.9.3 Data sets for Authority and Trust

- Kakol et al. (Further Reading 7) provides a manually annotated dataset that can be used for credibility prediction [48].
- Popat et al. (Further Reading 6) collected data from Wikipedia and snopes.com[49]

4.1.9.4 Further reading for Authority and Trust

1. Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
2. J. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, 46:604–632, 1999.
3. Yuting Liu , Bin Gao , Tie-Yan Liu , Ying Zhang , Zhiming Ma , Shuyuan He , Hang Li, BrowseRank: letting web users vote for page importance, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, July 20-24, 2008, Singapore, Singapore [doi>10.1145/1390334.1390412]
4. Gupta, Aditi, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. "Tweetcred: A real-time Web-based system for assessing credibility of content on Twitter." In Proc. 6th International Conference on Social Informatics (SocInfo). Barcelona, Spain. 2014.
5. Shafiza Mohd Shariff, Xiuzhen Zhang, Mark Sanderson. "On the credibility perception of news on Twitter: Readers, topics and features." Computers in Human Behavior 75 (2017) 785-794.
6. Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. "Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media." In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1003-1012. International World Wide Web Conferences Steering Committee, 2017.
7. Kakol, Michal, Radoslaw Nielek, and Adam Wierzbicki. "Understanding and predicting Web content credibility using the Content Credibility Corpus." Information Processing & Management 53, no. 5 (2017): 1043-1061.

4.1.10 Technicality

An article may be well written and grammatically understandable, but its content may cover concepts only understandable to people learned in a certain domain. These documents may deal with a technical issue or use a large proportion of technical terms.

4.1.10.1 Task for Technicality Measurement

For our Information Nutrition Label, we want to calculate a technicalness score, or technicality, for a document that indicates how hard it would be to understand for someone outside the field.

[48] https://github.com/s8811/reconcile-tags
[49] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/

4.1.10.2 Methods for Technicality Measurement

Similar to Readability, but more related to content than form, Technicality is a property of a document capturing the proportion of the domain-specific vocabulary used by the document. Style-based features are already captured by the readability score.

4.1.10.3 Data sets for Technicality Measurement

Data Sources:

- Terminology extraction software [50][51]
- Further tools are available from [52]
- In Wikipedia, external links provide a set of freely available tools under "Terminology Extraction"[53]
- Word frequency information[54] (English), in German[55][56][57], in other languages[58]

  Test cases and benchmarks:
- ACL RD-TEC[59]. QasemiZadeh, Behrang, and Anne-Kathrin Schumann. "The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods." In LREC. 2016.
- GENIA Corpus[60] is a popular corpus that has been used to evaluate various ATE algorithm for the last decade. In JATE2, instead of using the annotation file "GENIAcorpus302.xml", the 'concept.txt' containing a breakdown list of GENIA concepts and relations (more like ontology) are used as the "Gold Standard" (GS) list.

4.1.10.4 Further reading for Technicality Measurement

1. Justeson, John S., and Slava M. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text." Natural language engineering 1, no. 1 (1995): 9-27.
2. Dagan, Ido, and Ken Church. "Termight: Identifying and translating technical terminology." In Proceedings of the fourth conference on Applied natural language processing, pp. 34-40. Association for Computational Linguistics, 1994.
3. Pazienza, Maria, Marco Pennacchiotti, and Fabio Zanzotto. "Terminology extraction: an analysis of linguistic and statistical approaches." Knowledge mining (2005): 255-279.

4.1.11 Topicality

Topical documents are documents which cover topics that are in the current zeitgeist.

[50] https://github.com/termsuite/termsuite-core
[51] https://github.com/texta-tk/texta
[52] https://github.com/search?utf8=%E2%9C%93&q=terminology+extraction
[53] https://en.wikipedia.org/wiki/Terminology_extraction
[54] http://www.wordfrequency.info
[55] http://corpus.leeds.ac.uk/frqc/internet-de-forms.num
[56] http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html
[57] http://wortschatz.uni-leipzig.de/de/download
[58] https://web.archive.org/web/*/http://wortschatz.uni-leipzig.de/html/wliste.html
[59] https://github.com/languagerecipes/acl-rd-tec-2.0
[60] https://github.com/ziqizhang/jate/wiki/Evaluation-and-Dataset

4.1.11.1 Task for Topicality Detection

Topicality detection here means to decide whether the document is of current interest or not. One of the salient points of the negative effect of fake news was to falsely influence thinking about things in the current news cycle.

4.1.11.2 Methods for Topicality Detection

Extract the salient terms (keyterms) and entities of the document. Compare those terms to the terms found in recent news or publications, or search engine queries.

Tools:

- Text Mining Online[61]
- KeyPhrase Extraction[62][63]

4.1.11.3 Data sets for Topicality Detection

Current topics can be found on these sites, for example, ABC News[64], or lists of current events[65]. Current news and compiled multilingual lists of entities can be found at the UE-funded EMM NewsExplorer[66]

4.1.11.4 Further reading for Topicality Detection

1. Zafar, Muhammad Bilal, et al. Zafar, Muhammad Bilal, Parantapa Bhattacharya, Niloy Ganguly, Saptarshi Ghosh, and Krishna P. Gummadi. "On the Wisdom of Experts vs. Crowds: Discovering Trustworthy Topical News in Microblogs." In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 438-451. ACM, 2016
2. Wu, Baoning, Vinay Goel, and Brian D. Davison. "Topical trustrank: Using topicality to combat web spam." In Proceedings of the 15th international conference on World Wide Web, pp. 63-72. ACM, 2006.
3. Diakopoulos, Nicholas, and Arkaitz Zubiaga. "Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance." In ICWSM. 2014.

[61] http://textminingonline.com/how-to-use-stanford-named-entity-recognizer-ner-in-python-nltk-and-other-programming-languages Keyphrase extraction
[62] https://github.com/luffycodes/KeyphraseExtraction , https://github.com/Gelembjuk/keyphrases
[63] https://github.com/snkim/AutomaticKeyphraseExtraction
[64] http://abcnews.go.com/topics/
[65] http://libguides.umflint.edu/topics/current or http://www.libraryspot.com/features/currentevents.htm
[66] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

4.2 Rethinking Summarization and Storytelling for Modern Social Multimedia

Stevan Rudinac (University of Amsterdam, NL), Tat-Seng Chua (National University of Singapore, SG), Nicolas Diaz-Ferreyra (Universität Duisburg-Essen, DE), Gerald Friedland (University of California – Berkeley, US), Tatjana Gornostaja (tilde – Riga, LV), Benoit Huet (EURECOM – Sophia Antipolis, FR), Rianne Kaptein (Crunchr – Amsterdam, NL), Krister Lindén (University of Helsinki, FI), Marie-Francine Moens (KU Leuven, BE), Jaakko Peltonen (Aalto University, FI), Miriam Redi (NOKIA Bell Labs – Cambridge, GB), Markus Schedl (Universität Linz, AT), David Ayman Shamma (CWI – Amsterdam, NL), Alan Smeaton (Dublin City University, IE), and Lexing Xie (Australian National University – Canberra, AU)

Traditional summarization initiatives have been focused on specific types of documents such as articles, reviews, videos, image feeds, or tweets, a practice which may result in pigeonholing the summarization task in the surrounding of modern, content-rich multimedia collections. Consequently, much of the research to date has revolved around mostly toy problems in narrow domains and working on single-source media types. We argue that summarization and story generation systems need to refocus the problem space in order to meet the information needs in the age of user-generated content in different formats and languages. Here we create a framework for flexible multimedia storytelling. Narratives, stories, and summaries carry a set of challenges in big data and dynamic multi-source media that give rise to new research in spatial-temporal representation, viewpoint generation, and explanation.

4.2.1 Introduction

Social Multimedia [44] has been described as having three main components: content interaction between multimedia, social interaction around multimedia and social interaction captured in multimedia. Roughly speaking, this describes the interaction between traditional multimedia (photos and videos), mostly textual annotations on that media, and people interacting with that media. For almost a decade, fueled by the popularity of User-Generated Content (UGC), the bulk of research [18, 3, 40, 20, 1, 14, 9] has focused on meaningful extraction from any combination of these three points. With modern advancements in AI and computational resources [27, 19], we now realize that multimedia summarization and story telling has worked in isolated silos, depending on the application and media (object detection, video summarization, Twitter sentiment, etc.); a broader viewpoint on the whole summarisation and reduction process is needed. Consequently, this realization gives rise to a second set of research challenges moving forward. In this paper, we revisit and propose to reshape the future challenges in multimedia summarization to identify a set of goals, prerequisites, and guidelines to address future UGC. Specifically, we address the problems associated with increasingly heterogeneous collections both in terms of multiple media and mixed content in different formats and languages, the necessity and complexities of dense knowledge extraction, and the requirements needed for sense making and storytelling.

4.2.2 Related Work

Summarization problems. Content summarisation problems arise in different application domains, and are a long-standing interest of the natural language processing, computer vision, and multimedia research communities. Summarising long segments of text from a single or multiple documents is often done with extractive techniques, on which extensive surveys exist [28]. The problem of summarising image collections arises when there are e.g. large amounts of images from many users in a geographic area [36, 35], or about a particular social event [49], or when it is necessary to generate a summarizing description (caption) [9].

Similarly, it is often needed to shorten or find alternative presentation for long video sequences. Automatic story boards were probably first introduced by the CMU Informedia project [2] and video manga system is another early example of video-to-image summarisation in a comic-book layout [45]. Summarizing videos based on both audio and visual channels involved e.g. joint optimisation of cross-modal coherence [41], or matching of audio segments [16]. Summarization of ego-centric or surveillance videos attracted much attention recently, with the example approaches including finding important objects and actions [26] or constructing a map of key people in a known environment [50]. Many multimedia summarization problems are driven by real-world events at different time scales [49] and in the last decade there is also an increasing focus on large-scale social events reported online [16, 48]. This position paper examines the summarization problem more broadly, taking a step back from one particular media format to be summarized, and targeting a large range of applications.

Relevance criteria for summarization. Early approaches to information retrieval (IR) and summarization focused on relevance of the content presented to the user. However, by the end of 90s the community realized that users prefer diversified search results and summaries instead of results lists produced based on relevance criterion only [2]. While the application domains varied, since then most summarization approaches focused on finding a balance between relevance, representativeness and diversity. The Informedia project is one of the best known early examples following such paradigm in addressing, amongst others, the problem of video summarization [47]. However, as users may be more sensitive to irrelevant than (near) duplicate items, enforcing diversity without hurting relevance is very challenging. This is witnessed by a large body of research on e.g. image search diversification [17, 46, 23, 37, 15]. Social multimedia summarization has further found its way in diverse applications ranging from personalized tweet summarization [34] to visual summarization of geographic areas and tourist routes [36, 35, 12, 17] for POI recommendation and exploration. With the increased availability of affordable wearables, in recent years lifelogging has started gaining popularity, where the goal is to generate a diary or a record of the day's activities and happenings by creating a summary or a story from the video/image data gathered [11, 22]. Progress has been made in summarizing heterogeneous user-generated content with regards to relevance, representativeness, and diversity [2]. However, relevance criteria and their interplay may be much more complex than commonly assumed [36] and, in case of visual content, include additional factors such as content popularity, aesthetic appeal and sentiment. Thus we call for rethinking the foundations of summarization and storytelling.

Benchmarks and Formalization Efforts. For almost two decades, common datasets, tasks, and international benchmarks fueld research on summarization and storytelling [13, 5, 31]. A typical task involved automatically generating a shorter (e.g. 100-word) summary of a set of news articles. TRECVID BBC Rushes summarization was probably the first systematic effort in the multimedia and computer vision communities focusing on video summarization [30]. The task involved reducing a raw and unstructured video captured during the recording of a

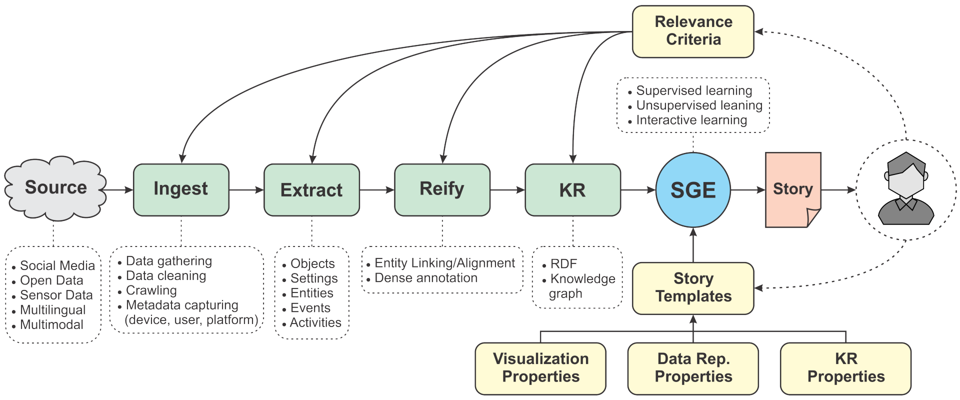

Figure 2 Pipeline of our proposed framework for generating narratives, stories and summaries from heterogeneous collections of user generated content and beyond.

TV series to a short segment of just a couple of minutes. Another well-known example is the ImageCLEF 2009 Photo Task, which revolved around image diversification [23]. The participants were expected to produces image search results covering multiple aspects of a news story, such as the images of "Hillary Clinton", "Obama Clinton" and "Bill Clinton" for a query "Clinton". Image search diversification has also been a topic of an ongoing MediaEval Diverse Social Images Task, run annually since 2013 [15].

Although many people intuitively understand the concept of summarization, the complexity of the problem is best illustrated by the difficulties in even unequivocally defining a summary [33]. So, instead of focusing on strict definitions, most benchmarks took a pragmatic approach by conducting either intrinsic or extrinsic evaluation [13]. In intrinsic evaluation an automatically generated summary is compared directly against a "standard", such as summaries created by the humans. On the other hand, extrinsic evaluation measures the effectiveness of a summary in fulfilling a particular task as compared with the original set of documents (e.g. text, images or videos). Over the years many interesting metrics for evaluating (text) summaries were proposed, such as recall-oriented understudy for gisting evaluation (ROUGE) [25], bilingual evaluation understudy (BLEU) [32] and Pyramid Score [29]. Some of these were later on successfully adapted to the visual domain [24, 36]. These initiatives had an impact on the progress in the field of summarization. However, their almost exclusive focus on a single modality (e.g. text or visual) or language and the traditional tasks (e.g. text document and video summarization or search diversification) does not reflect the richness of social multimedia and the complex use cases it brings.

4.2.3 Framework Overview

First, we take a step back and look at a media-agnostic birds-eye view of the problem (see Figure 2). We therefore imagine a generic framework that follows the requirements as driven by the user, instead of the technology. Figure 2 shows an overview of the concept, which follows the standard pattern of a media pipeline along the "ingest," "extract," "reify" paradigm. The goal of the framework is to create a story for the user, who is querying for information using a set of relevance criteria. Before doing that, we assume the user has configured the framework somehow, e.g. to choose some visualization template and define basic properties of the story. We then assume a tool that would query a set of sources from the Internet or elsewhere, download ("ingest") the data, "extract" relevant information and then "reify" it in a way that it can be added into some standardized Knowledge Representation (KR). The knowledge representation would then, in connection with the initial configuration, be used to create the final story. We will next discuss technical and other challenges to be addressed by the community in order to put flesh onto our bare bones framework.

4.2.4 Challenges and Example Application Domains

A framework for holistic storytelling brings a new set of research challenges and also reshapes some of the more traditional challenges in UGC. We identify these as *storytelling challenges* which include handling of time/temporality/history, dynamic labeling of noise, focused story generation, tailoring to impartiality or a viewpoint, quality assessment and explainability as well as *UGC challenges* which include ethical use, multi source fusion, multilinguality and multimodality, information extraction, knowledge update and addition of new knowledge, staying agnostic to specific application, supporting various types of open data, portability and finding a balance between depth and breadth. We now describe a set of application domains that illustrate some of the aforementioned challenges.

Smart urban spaces. Increased availability of open data and social multimedia has resulted in the birth of urban computing [51] and created new possibilities for better understanding a city. Although spontaneously captured, social multimedia may provide valuable insights about geographic city regions and their inhabitants. For example, user-generated content has been used to create summaries of geographic areas and tourist routes in location recommendation systems [35, 36]. Sentiment extracted from social multimedia, in combination with neighborhood statistics was also proven invaluable for a more timely estimation of city livability and its causes [8]. Similarly, when looking for signs of issues such as neighborhood decay or suboptimal infrastructure, city administrators are increasingly monitoring diverse UGC streams, ranging from social media and designated neighborhood apps to data collected by mobile towers and wearables. Efficient approaches to summarization and storytelling are needed to facilitate exploration in such large and heterogeneous collections.

Field study/survey. Consumer-produced multimedia contains data which is not only relevant to the reason for creating and sharing it but also for other applications. As a side effect this information could be used for field studies of other kinds, if it can be retrieved in a timely fashion. The framework we propose and especially the kind of tools that it leads to should enable empirical scientists of many disciplines to leverage this data for field studies based on extracting required information from huge datasets. This currently constitutes a gap between the elements of what multimedia researchers have shown is possible to do with consumer-produced big data and the follow-through of creating a comprehensive field study framework supporting scientists across other disciplines. To bravely bridge this gap, we must meet several challenges. For example, the framework must handle unlabeled and noisily labeled data to produce an altered dataset for a scientist — who naturally wants it to be both as large and as clean as possible. We must also design an interface that will be intuitive and yet enable complex search queries that rely on feature and statistics generation at a large scale.

Business intelligence. User generated content is a valuable source of information for companies and institutions. Business information can be obtained by analyzing what the public is saying about a company, its products, marketing campaigns and competitors. Traditionally business intelligence relied on facts and figures collected from within the organisation, or provided by third-party reports and surveys. Instead of surveys, direct feedback can be obtained by listening to what people are saying on Social Media, either directed at their own social circle, or directly at the company in the case of web care conversations. Content can consist of textual messages or videos, for example product reviews. Besides the volume of messages, the sentiment of messages is important to analyze into positive and negative aspects. The amount of user generated content can easily add up to thousands of messages on a single topic, so summarization techniques are needed to efficiently process the wealth of information available [6].

Health and Wellness. There is a wealth of data about our health and wellness which is generated digitally on an individual basis. This includes genomic information from companies like 23andme[67] which uses tissue samples from individuals to generate information about our ancestry as well as about our possible susceptibility to a range of inherited diseases. We also have information about our lifestyles which can be gathered from our social media profiles and information about our physical activity levels and sports participation from any fitness trackers that we might wear or use. When we have health tests or screening we can have indications of biomarkers from our clinical tests for such things as cholesterol levels, glucose levels, etc. We have occasional once-off readings of our physiological status via heart and respiration rates and increasingly we can use wearable sensors to continuously monitor glucose, heart rate etc. to see how these change over time. From all of this personal sensor data there is a need to generate the *"story of me"*, telling my health professional and me how well or healthy I am now, whether my health and wellness is improving or is on the slide, and if there's anything derived from those trends that I should know.

Lifelogs. In this use case a large amount of first-person ethnographic video or images taken from a wearable camera over an extended period of days, weeks, months or even years, has been generated. Such a collection may be augmented and aligned with sensor data such as GPS location or biometric data like heart rate, activity levels from wearable accelerometers or stress levels from galvanic skin response sensors. There is a need to summarize each day's or week's activities to allow reviewing or perhaps to support search or browsing through the lifelog. Summaries should be visual, basically selecting a subset of images of videos, and applications could be in memory support where a summary of a day can be used to trigger memory recall [7]. In this case the visual summary should incorporate events, objects or activities which are unusual or rare throughout the lifelog in preference to those which are mundane or routine like mealtimes, watching TV or reading a newspaper which might be done every day [21].

Entertainment. Multimedia summarization and storytelling can also serve to fulfill a pure entertainment need. Respective approaches could, for instance, support event-based creation of videos from pictures and video clips recorded on smart phones. To this end, they would automatically organize and structure such user-generated multimedia content, possibly in low quality, and subsequently determine the most interesting and suited parts in order to tell the story of a particular event, e.g., a wedding. The multimedia material considered by such an event-based storytelling approach is not necessarily restricted to a single user, but could automatically determine and select the best images / scenes from the whole audience at the event, or at least those choosing to share material.

[67] http://www.23andme.com

4.2.5 Use Cases

When rethinking the requirements, we primarily analyzed two types of use cases: summarization and storyboarding. Summarization has traditionally involved document summarization, i.e. reducing one or more pieces of text into a shorter version and video summarization where multiple or long videos are reduced to a shorter version. Summaries could include an abridged report of an event or a how-to instruction with the main points to perform a task. As data is increasingly available in many modalities and languages, it is possible to generate a summary from and in multiple modalities and languages according to the user's information request. Large events such as elections or important sports competitions are covered by many channels, including traditional media and different Social Media including text messages, images and video. New directions for summarization include interactive summaries of UGC opinions or sentiment-based data visualization, and forecasting including prediction of electoral results, product sales, stock market movements and influenza incidence [39].

While users of music streaming services are often drowning in the number of music pieces to choose from, getting an overview of a certain music genre or style, which serves an educational purpose, is barely feasible with current recommender systems technology. Addressing this, we need algorithms that automatically create consistent and entertaining summaries of the important songs of a given genre or style considering the genre's evolution over time. Such approaches need to identify the most representative and important music, use automatic structural music segmentation techniques [43, 10, 4], decide on the most salient parts, and present them in a way that connects them. Ideally, the approaches should also consider cultural perspectives to take into account that the meaning of genres such as folk music may change depending on the cultural background of the listener such as the country, among other aspects [38].

A storyboard is a summary that conveys a change over time. This may include a recount of the given input in order to tell an unbiased story of an event, e.g. the Fall of the Berlin wall or the Kennedy murder. It may also aim to present or select facts to persuade a user to perform a particular action or change opinion, e.g. pointing out the likely murderer in the Kennedy case. If the input is open-ended, the summary may be structured by background information, e.g. a composite clip giving a visual summarization of an event (such as a concert, a sports match, etc.) where the summarized input is provided by those attending the event but the story is structured according to a timeline given by background information.

4.2.6 Prerequisites

Once user generated content has been gathered, extracted, and reified, it should be expressed in a KR. This is a step prior to the generation of stories and summaries which aims to describe the information of interest following a representation formalism. Some of the knowledge representation formalisms widely adopted in the multimedia community are Resource Description Framework (RDF) and Knowledge Graph. The selection of one approach over the others is tightly connected with the purpose of the summary/story and the technique used for its construction. This means that knowledge must be represented using a language with which the Story Generation Engine can reason in order to satisfy complex relevance criteria and visualization requirements (templates) specified by the users. These relevance criteria and visualization requirements imply a set of desired properties on the data and KR, as well as the end result presented to the user, which are fundamental for summarization and storytelling.

Table 2 Properties data representation should have for facilitating effective summarization and storytelling.

Data Representation Properties

Location	*Time*	*Observed*
Single ⇌ Distributed	Scheduled ⇌ Unplanned	Entity-driven ⇌ Latent
Physical ⇌ Virtual	Short ⇌ Long	
Personal ⇌ Public/Shared	Recurrent ⇌ One-off	
Independent ⇌ Cascaded		

4.2.6.1 Representation Properties

Complex user information needs and the relevance criteria stemming from them require novel (multimodal and multilingual) data representations. In Table 2 we list some critical prerequisites they should fulfill.

Time: The "events" described by a story could have very different properties. For example, an event could be *scheduled* (e.g. Olympic Games) or *unplanned* (e.g. a terrorist attack). In the former case relevance criteria and the visualization templates could be easier to foresee, but an effective data representation should accommodate the latter use case as well. Similarly, the events could have a *longer* (e.g., studies abroad) or *shorter* (e.g., birthday) duration. Finally, data representation should ideally accommodate both *recurrent* and *once off* events.

Location: Although multimedia analysis has found its way in modeling different aspects of geographic locations [15, 35, 42], most related work addressed specific use cases and little effort has been made in identifying general "spatial" criteria underlying data representations should satisfy. In this regard, the representation should account for the events occurring at a *single* (e.g. rock concert) or *distributed* location (e.g. Olympic Games). In both cases those locations can be further *physical* or *virtual*. On the other hand, the events of interest can be *personal* or *public/shared*. While in the former case the content interpretation and relevance criteria may have a meaning for a particular individual only, the later is usually easier to analyze due to a higher "inter-user agreement". Finally, data representation should be designed with the awareness that the aforementioned types of events could additionally be *independent* or *cascaded*.

Observed: In many analytic scenarios the summaries and stories presented to the user contain well-defined named *entities*, i.e. topics, people, places and organizations. An example would be a well-structured news article covering a political event. Yet the topics of interest may be *latent*, which is particularly common in social media discussions. For example, a public servant sifting through millions of social media posts in an attempt to verify an outbreak of a new virus may be interested in various unforeseen and seemingly unrelated conversations, which together provide conclusive evidence. Therefore, a good data representation should ideally provide support for both.

4.2.6.2 Representation Properties

Building on best practices from the semantic web community, the results of ingestion, extraction and reification (cf. Figure 2) should be further organized in a knowledge representation. Example candidates include RDFs and knowledge graph. The KR should be flexible enough to allow for *temporal*, *spatial* and *observed* properties of the events discussed in subsection 4.2.6.1. It should further support both *implicit* and *explicit* relations between the items, as well as

Table 3 Properties a knowledge representation should have.

Knowledge Representation Properties	
Implicit ⇌ Explicit	Independent ⇌ Correlated/Casual
Uniqueness/Representativeness	Support for different semantic levels

Table 4 Story properties that should be facilitated by the story generator engine.

Story Properties	
Functional ⇌ Quality	Modality-preserving ⇌ Cross-modal
Self-contained ⇌ Stepping-stone	Static ⇌ Dynamic/Interactive
Succinct ⇌ Narrative	Factual ⇌ Stylistic
Abstractive ⇌ Generative	Generic ⇌ Personalized

their modification "on the fly" (cf. Table 3). The events and their building blocks could further be *independent* and *correlated/casual*. To facilitate a wide range of possible relevance criteria as well as their complex interplay, the KR should also include notions of importance, representativeness and frequency. Finally, the content interpretation and user information needs can be specified at different semantic levels, which in case of multimedia range from e.g. term or pixel statistics, semantic concepts, events and actions to the level of semantic theme and complex human interpretations involving aesthetic appeal and sentiment. Supporting a wide range of relevance is therefore a necessary condition for facilitating creation of effective and engaging summaries and stories.

4.2.6.3 Properties

Given the content, data and KRs and the user information needs, the output of the pipeline depicted in Figure 2 is the story (or summary) presented to the user. Below we enumerate a number of criteria an ideal set of "story templates" should satisfy (see Table 4). A story should satisfy both *functional* (e.g. fulfilling a purpose) and *quality* (e.g. metrical) requirements [13]. The importance of a particular requirement should ideally be learned from user interactions. The system should further support *self-contained/interpretable* and *stepping-stone/connector* type of summaries. While the former by itself provides an insight into a larger multimedia item or a collection, the later serves a goal further on the horizon, such as faster collection exploration. Additionally, the design should accommodate both *succinct* and *narrative*, as well as *abstractive* and *generative* stories. With regard to the input and output modalities and languages, support should be provided for *modality-preserving* and *cross-modal* and/or *cross-lingual* use-cases. In many scenarios, user information needs can be satisfied with a *static* story. However, the size and heterogeneity of a UGC collection as well as the complexity of user information needs make *interactive* summarization and storytelling increasingly popular. Depending on the information needs, a *factual* or *stylistic* summary may be desirable, which is why the system should support both flavors and perhaps allow for interactive learning of their balance. Finally, while a *generic* story may be sufficient for some, *personalization* should also be supported.

4.2.7 Conclusion

Motivated by an observation about discrepancies between state of the art research on the one hand and the increasing richness of user generated content and the accompanying complex user information needs on the other, we revisit the requirements for multimedia summarization and storytelling. We reiterate the importance of summarization and storytelling for facilitating efficient and appealing access to large collections of social multimedia and interaction with them. Our proposed framework identifies a set of challenges and prerequisites related to data and KR as well as the process of their creation, i.e. ingestion, extraction and reification. We further make an inventory of the desirable properties a story should have for addressing a wide range of user information needs. Finally, we showcase a number of application domains and use cases that could serve as the catalyst for future research on the topic.

References

1 J. Bian, Y. Yang, H. Zhang, and T. S. Chua. Multimedia summarization for social events in microblog stream. *IEEE Trans. Multimedia*, 17(2):216–228, Feb 2015.

2 Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR '98*, pages 335–336, New York, NY, USA, 1998. ACM.

3 Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *ACM IMC '07*, pages 1–14, 2007.

4 W. Chai. Semantic segmentation and summarization of music. *IEEE Signal Process. Mag.*, 23(2), 2006.

5 Hoa Trang Dang. Overview of DUC 2006. In *DUC '06*, 2006.

6 Lipika Dey, Sk Mirajul Haque, Arpit Khurdiya, and Gautam Shroff. Acquiring competitive intelligence from social media. In *MOCR AND '11*, page 3. ACM, 2011.

7 Aiden R. Doherty, Steve E. Hodges, Abby C. King, Alan F. Smeaton, Emma Berry, Chris J.A. Moulin, Siân Lindley, Paul Kelly, and Charlie Foster. Wearable cameras in health. *Am J Prev Med*, 44:320–323, March 2013.

8 Joost Boonzajer Flaes, Stevan Rudinac, and Marcel Worring. What multimedia sentiment analysis says about city liveability. In *ECIR '16*, pages 824–829, 2016.

9 Tatjana Gornostay (Gornostaja) and Ahmet Aker. Development and implementation of multilingual object type toponym-referenced text corpora for optimizing automatic image description generation. In *Dialogue '09*, 2009.

10 Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Trans. Audio, Speech, Language Process.*, 14(5):1783–1794, Sept 2006.

11 Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. Lifelogging: Personal big data. *Found. Trends Inf. Retr.*, 8(1):1–125, June 2014.

12 Qiang Hao, Rui Cai, Xin-Jing Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. Generating location overviews with images and tags by mining user-generated travelogues. In *ACM MM '09*, pages 801–804, New York, NY, USA, 2009. ACM.

13 Donna Harman and Paul Over. The DUC summarization evaluations. In *HLT '02*, pages 44–51, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

14 Richang Hong, Jinhui Tang, Hung-Khoon Tan, Chong-Wah Ngo, Shuicheng Yan, and Tat-Seng Chua. Beyond search: Event-driven summarization for web videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7(4):35:1–35:18, December 2011.

15 Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. Result diversification in social image retrieval: a benchmarking framework. *Multimed Tools Appl*, 75(2):1301–1331, Jan 2016.

16 Lyndon Kennedy and Mor Naaman. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *WWW '09*, pages 311–320. ACM, 2009.

17 Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *ACM WWW '08*, pages 297–306, 2008.

18 Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. *Twitter Sentiment Analysis: The Good the Bad and the OMG!*, pages 538–541. AAAI Press, 2011.

19 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *NIPS '12*, pages 1097–1105. Curran Associates, Inc., 2012.

20 Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *ACM WWW '10*, pages 591–600, 2010.

21 Hyowon Lee, Alan F Smeaton, Noel E O'Connor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin. Constructing a SenseCam visual diary as a media process. *Multimedia Syst*, 14(6):341–349, 2008.

22 Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE CVPR '12*, pages 1346–1353, June 2012.

23 Monica Lestari Paramita, Mark Sanderson, and Paul Clough. *Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009*, pages 45–59. 2010.

24 Yingbo Li and Bernard Merialdo. Vert: Automatic evaluation of video summaries. In *ACM MM '10*, pages 851–854, 2010.

25 Chin Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL '04 Workshop*, pages 74–81, 2004.

26 Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE CVPR '13*, pages 2714–2721, 2013.

27 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS '13*, pages 3111–3119. Curran Associates, Inc., 2013.

28 Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. *Mining text data*, pages 43–76, 2012.

29 Ani Nenkova and Rebecca J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152, 2004.

30 Paul Over, Alan F. Smeaton, and George Awad. The TRECVid 2008 BBC rushes summarization evaluation. In *ACM TVS '08*, pages 1–20, 2008.

31 Karolina Owczarzak and Hoa Trang Dang. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *TAC '11*, 2011.

32 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL '02*, pages 311–318, 2002.

33 Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Comput. Linguist.*, 28(4):399–408, December 2002.

34 Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. Personalized time-aware tweets summarization. In *ACM SIGIR '13*, pages 513–522, 2013.

35 S. Rudinac, A. Hanjalic, and M. Larson. Generating visual summaries of geographic areas using community-contributed images. *IEEE Trans. Multimedia*, 15(4):921–932, June 2013.

36 S. Rudinac, M. Larson, and A. Hanjalic. Learning crowdsourced user preferences for visual summarization of image collections. *IEEE Trans. Multimedia*, 15(6):1231–1243, Oct 2013.

37 Mark Sanderson, Jiayu Tang, Thomas Arni, and Paul Clough. *What Else Is There? Search Diversity Examined*, pages 562–569. 2009.

38 Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *J Intell Inf Syst*, 41:523–539, December 2013.

39 Harald Schoen, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, and Peter Gloor. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013.

40 David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. Tweet the debates: Understanding community annotation of uncollected sources. In *ACM WSM '09*, pages 3–10, 2009.

41 Hari Sundaram, Lexing Xie, and Shih-Fu Chang. A utility framework for the automatic generation of audio-visual skims. In *ACM MM '02*, pages 189–198. ACM, 2002.

42 Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016.

43 Mi Tian and Mark B. Sandler. Towards music structural segmentation across genres: Features, structural hypotheses, and annotation principles. *ACM Trans. Intell. Syst. Technol.*, 8(2):23:1–23:19, October 2016.

44 Yonghong Tian, Jaideep Srivastava, Tiejun Huang, and Noshir Contractor. Social multimedia computing. *Computer*, 43(8):27–36, August 2010.

45 Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. Video manga: generating semantically meaningful video summaries. In *ACM MM '99*, pages 383–392. ACM, 1999.

46 Reinier H. van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. Visual diversification of image search results. In *ACM WWW '09*, pages 341–350, 2009.

47 H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens. Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52, May 1996.

48 Lexing Xie, Apostol Natsev, John R Kender, Matthew Hill, and John R Smith. Visual memes in social media: tracking real-world news in youtube videos. In *ACM MM '11*, pages 53–62. ACM, 2011.

49 Lexing Xie, Hari Sundaram, and Murray Campbell. Event mining in multimedia streams. *Proc. IEEE*, 96(4):623–647, 2008.

50 Shoou-I Yu, Yi Yang, and Alexander Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *IEEE CVPR '13*, pages 3714–3720, 2013.

51 Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):38:1–38:55, September 2014.

52 Abdel Karim Al Tamimi, Manar Jaradat, Nuha Al-Jarrah, and Sahar Ghanem. Aari: automatic arabic readability index. *Int. Arab J. Inf. Technol.*, 11(4):370–378, 2014.

53 Maik Anderka. *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia.* Dissertation, Bauhaus-Universität Weimar, June 2013.

54 Maik Anderka, Benno Stein, and Nedim Lipka. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 981–990. ACM, August 2012.

55 Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742–753, 2012.

56 Michael L Bernard, Barbara S Chaparro, Melissa M Mills, and Charles G Halcomb. Comparing the effects of text size and format on the readibility of computer-displayed times new roman and arial text. *International Journal of Human-Computer Studies*, 59(6):823–835, 2003.

57 Bartosz Broda, Bartlomiej Niton, Wlodzimierz Gruszczynski, and Maciej Ogrodniczuk. Measuring readability of polish texts: Baseline experiments. In *LREC*, pages 573–580, 2014.

58 Kevyn Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.

59 Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493, 2008.

60 Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493, 2008.

61 Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325, 2014.

62 Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

63 Thomas François. An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107. Linköping University Electronic Press, 2014.

64 Richard H Hall and Patrick Hanna. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23(3):183–195, 2004.

65 Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. Event detection and factuality assessment with non-expert supervision. 2015.

66 Mary Levis, Markus Helfert, and Malcolm Brady. Information quality management: Review of an evolving research area. 01 2007.

67 Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the tenth conference on computational natural language learning*, pages 109–116. Association for Computational Linguistics, 2006.

68 Amnon Lotan, Asher Stern, and Ido Dagan. Truthteller: Annotating predicate truth. 2013.

69 A-L Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, MGJ van Erp, AM Schoen, CM van Son, et al. Meantime, the newsreader multilingual event and time corpus. 2016.

70 Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics, 2008.

71 Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *CoRR*, abs/1702.05638, 2017.

72 Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659, 2013.

73 Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227, 2009.

74 Roser Saurí and James Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.

75 Sanja Stajner and Horacio Saggion. Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *IJCNLP*, pages 374–382, 2013.

76 Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 352–357, 2017.

77 Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4):217–248, 2013.

78 Richard Y. Wang and Diane M. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

79 Beverly L Zakaluk and S Jay Samuels. *Readability: Its Past, Present, and Future.* ERIC, 1988.

80 M. Potthast, S. Köpsel, B. Stein, and M. Hagen. Clickbait Detection. In Nicola Ferro et al., editors, *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 16)*, volume 9626 of *Lecture Notes in Computer Science*, pages 810–817, Berlin Heidelberg New York, March 2016. Springer.

81 S. Heindorf, M. Potthast, B. Stein, and G. Engels. Vandalism Detection in Wikidata. In Snehasis Mukhopadhyay et al., editors, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 16)*, pages 327–336. ACM, October 2016.

82 H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, and B. Stein. Building an Argument Search Engine for the Web. In *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 17)*, EMNLP 2017, Kopenhagen, September 2017.

Participants

- Tat-Seng Chua
National University of Singapore, SG
- Nicolas Diaz-Ferreyra
Universität Duisburg-Essen, DE
- Gerald Friedland
University of California – Berkeley, US
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Anastasia Giachanou
University of Lugano, CH
- Tatjana Gornostaja
tilde – Riga, LV
- Gregory Grefenstette
IHMC – Paris, FR
- Iryna Gurevych
TU Darmstadt, DE
- Andreas Hanselowski
TU Darmstadt, DE
- Xiangnan He
National University of Singapore, SG
- Benoit Huet
EURECOM – Sophia Antipolis, FR
- Kalervo Järvelin
University of Tampere, FI
- Rosie Jones
Microsoft New England R&D Center – Cambridge, US
- Rianne Kaptein
Crunchr – Amsterdam, NL
- Krister Lindén
University of Helsinki, FI
- Yiqun Liu
Tsinghua University – Beijing, CN
- Marie-Francine Moens
KU Leuven, BE
- Josiane Mothe
University of Toulouse, FR
- Wolfgang Nejdl
Leibniz Universität Hannover, DE
- Jaakko Peltonen
Aalto University, FI
- Isabella Peters
ZBW – Dt. Zentralbib. Wirtschaftswissenschaften, DE
- Miriam Redi
NOKIA Bell Labs – Cambridge, GB
- Stevan Rudinac
University of Amsterdam, NL
- Markus Schedl
Universität Linz, AT
- David Ayman Shamma
CWI – Amsterdam, NL
- Alan Smeaton
Dublin City University, IE
- Benno Stein
Bauhaus-Universität Weimar, DE
- Lexing Xie
Australian National University – Canberra, AU