LANGUAGE INTELLIGENCE @WORK

www.liatwork.com

News, Disinformation & Language Intelligence

Orientation Paper

How Language Intelligence can combat disinformation and support reliable journalism for the News Media

LT-INNOVATE.ORG

SAILLABS TECHNOLOGY

Contents

Executive Summary

"Fake News" or better, disinformation, is false information deliberately and often covertly spread (as by the planting of rumours) in order to influence public opinion or obscure the truth[1] . The ever-increasing channels on the Internet support communications of any kind. These zetta-bytes of information make the detection of disinformation a super-human task. Artificial Intelligence together with Language Technologies (referred to as "Language Intelligence" or LI by LT-Innovate) can effectively scale up the journalistic process of discovery, source identification, and verification. This refers to (multilingual) text as well as images/videos and sound. Currently, large publishers or news agencies like Associated Press or New York Times are experimenting with automated processes. Challenges regarding automated fact checkers include:

- ➔ Large amounts of (relevant) data are necessary which tend to slow down processes
- ➔ Despite enormous advances in neural networks/algorithms, automated contextualization is still a problem
- ➔ Multilingual sources are necessary for serious journalism but different language components are at different levels of readiness for automation

On the other hand, LI can support the News Media at other levels than disinformation detection, e.g. data enhancement (to exploit news archives and help machine-learning for automated processes), cross-media search (through speech to text and vice versa), or new delivery channels for new business models (pay-per-read, personalisation).

In November 2018 SAIL LABS and LT-Innovate jointly organised a conference targeting Fake News and associated phenomena. The event took place in Vienna and hosted a wide array of international attendees. This Orientation Paper is the result of that conference as well as of interviews and desk research (see Annex for more details on method and sources).

Foreword

"Fake news" is not a new phenomenon. Disinformation has always been used to dethrone potentates or start wars. What is new in the 21st century is the spread of digital communication systems that reach out to mass consumers in a border- and limitless fashion. Democratic processes such as elections are a primary target but they are just the tip of the democracy-threatening iceberg that disinformation represents. At EU level, disinformation is a top priority since the East Stratcom Task Force against Russian disinformation was created in 2015. Cross-border initiatives are necessary to combat the danger of disinformation without censuring free speech.

New technologies can support this endeavour, in particular artificial intelligence paired with language technologies provide a promising mix to tackle this enormous challenge!
The amount of data and the speed of their transmission requires the use of automated processes and algorithm to detect and analyse disinformation in real time.

[1] Merriam-Webster

This "Orientation Paper" represents the view of language intelligence tools developers, researchers, media representatives and NGOs that are all in one way or another affected by "fake news". It covers available solutions & initiatives and identifies what is needed to be done in the near future to guarantee a free and serious News Media landscape in Europe.

1. News, language and automation

Figure 1 Source: Presentation P. Cochrane

News means information about recent events that embodies new and noteworthy content. News is expressed in human – written or spoken – language, often combined with illustrative media such as photos or videos.
In addition, the entities reported - persons, organisations, actions, statements, places, times, outcomes - are verifiable, which means that unlike stories or jokes, news content can be evaluated according to real-world truth conditions. However, this does not prevent untrue or partially true (i.e. fake) news stories from circulating unimpeded.

Large, medium and small corporates in almost any sector, as well as government agencies, NGOs and similar, are used to monitoring news about themselves as found on major networks and social media. They also increasingly generate news stories or engage in fact-based journalism (text, video) about their own activities that target their own clients and stakeholders.

The news cycle is now 24/7, making it vital for news production offices and monitoring agencies to operate and respond to news content in real time around the clock. This increases the need for productivity tools to help create, evaluate and react to the global content flow of information in multiple languages, putting added time pressure on newsroom capacity and increasing the need to innovate and automate.

Very short-form info bites - such as tweets - usually depend for context on either physically surrounding tweets in the stream or on a reader's implicit background knowledge. In the latter case, the lack of such background can lead to easy misreadings of the true intent or value of a given news tweet. Hence the danger of apparently context-free content in a rapid news cycle. Putting content into better and broader context will increasingly

require a more intelligent, "conversational" approach.

Language in its written, spoken and visual form is, of course, the key medium of news content. The tools and technologies that have been developed to create and manage this content can therefore be called collectively "**language intelligence** (LI)", a term that may be applied in a more general context and which is not confined to the media alone. The application of LI provides a critical mechanism across the digital world, where speed, data volume, trust, cost and quality of knowledge all need to be carefully handled and balanced to deliver reliable products to networked customers on a global scale. As well as opening up new opportunities in news production and delivery, LI can play a central role in the constant fight against disinformation.

This report attempts to characterize the various roles that language intelligence plays in the News value chain by asking the following three questions and providing answers to them:

- **Are there reliable automatic decision procedures for identifying the truth/falsity of news and similar content in general or in specific cases? Problem solution technologies.**

- **Which available LI technologies can help news producers most to accelerate processes while lowering risk and complexity? Technologies along the News value chain.**

- **What kinds of language solutions are needed and likely to emerge? The way forward.**

Target readers of this report:

1. News media and others (NGOs, public and private sector) that produce or consume news, or monitor their presence in news channels (print, online, broadcast).

2. LI developers: The news value chain is in dire need of technologies that support the creation and marketing of reliable news. The report can help adapt existing tools for the news sector or develop targeted tools for these needs.

3. Research (private and public): The report can act as an indicator for research to engage in developments that target the solution of real problems of the sector not met so far.

4. Public sector: Some problems are the result of societal factors and (a lack of) media literacy. The public sector can help to engage in activities that lead to a propitious environment for the news media and that foster media literacy in education.

2. Disinformation and the fake news challenge

Disinformation is a social condition, like crime, not a plumbing problem that you can fix easily.
Vincent Tripodi

News content forms part of democracy and the open economy, providing checkable information about day-to-day events in a given society or geography. Today this content is under pressure from two directions

a) the expansion of global social media that enable "fake news" or disinformation to be created and circulated as easily as truthful news, and
b) the power of artificial intelligence applications to simultaneously accelerate and multiply the production of news content of all kinds, including fake news.

What is "fake news"? A brief look at terminology to set a level playing field for this report.

The issue of definition was raised in the Ethical Working Group (Working Group 1 of the conference). While there might be many definitions of "fake news", this report will use the term "disinformation" to refer to "false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit. It is driven by the production and promotion of disinformation for economic gains or for political or ideological goals"[2] . One reason why the High level Expert Group on Disinformation rejected the term Fake News was that often, fabricated news with the intention to deceive takes a "blended" approach and uses – contorted – facts.

This toxic form of content potentially causes social disturbance, political and territorial insecurity, and commercial damage. We will collectively refer to its various forms as disinformation. It is most egregiously expressed digitally as textual, image or video content that *intentionally* conveys erroneous information virally about some entity, relationship or event. But in fact, there is a spectrum of disinformation, ranging from the outright intention to dis-inform to the use of sensationalist language to act as clickbait on an online information source to attract more viewers and monetize the site by attracting advertisers.

"Misinformation" to the contrary is usually used for unintentional "fake" news, e.g. due to sloppy sources or editing but without a hidden agenda.

The disinformation problematic :

It all starts in the minds and hearts of people who are influenced by misinformation (K. Varga, Bakamo). Apart from fake news creation for pecuniary reasons (e.g. to boost enormous click rates - so-called clickbaits), disinformation holds emotional content that readers are more likely to share as they feel attracted and understood. An example of information sources of the French elections showed the following pattern (Source: K. Varga):

- 50% of information came from traditional media
- 7% from electoral campaigns
- 20% from extended sources: civil society, investigative journalism (also ad hearing to journalistic standards)

[2] Final report of the High Level Expert Group on Fake News and Online Disinformation, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

- 19% reframe sections of traditional media so that it serves their purpose (political, social etc.)
- 4% Alternative: conspiracy theories etc.

It did not come as a surprise that those messages which were shared, re-tweeted etc. were the reframed and alternative messages. People on the Internet crave to "belong" to a community (no matter which one), identify themselves with it (even though it might be a negative identification), and feel valued or important.

The psychological problem goes one step further: If people define situations as real, they *are real in their consequences* (G. Czech, Austrian Red Cross). If lies are repeated often enough, people will believe them (P. Cochrane, University of Suffolk). The "post-truth" society needs effective education in "media literacy" (C. Gsodam) that should be encouraged by individual countries as well as the EU. Already, though, multilingual AI technologies ("language intelligence") can act in the front line to support the human battle against zetta-amounts of information sources of data that need to be sifted through to check their reliability.

3. Language Intelligence and Journalism

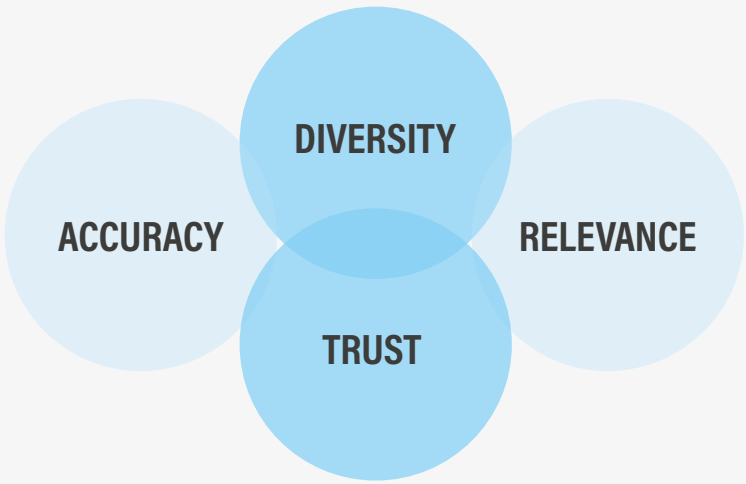


Figure 2: Basics of serious journalism where AI/LI steps in

LI covers all technologies that process language in order to augment cognitive linguistic tasks typically carried out by humans. These include speaking, reading, writing, searching/finding, editing, summarizing, translating, paraphrasing, and the like.

The primary purpose of LI is to simplify, accelerate and interconnect those actions and processes that can be automated, thereby freeing up more time and effort for the central tasks involving human discernment, intuition, imagination and investigative journalism.

ACCURACY: In the journalistic Code of Conducts, accuracy is key to serious journalism. It is the "Veracity, accu-

racy, credibility" testing that takes up enormous amounts of time to maintain a high level of reporting. Regarding automated image recognition, from an editorial perspective, journalists are not keen on the idea that an image is 70% correct (although this is an enormous achievement of the technology!). There is still a long way to go until journalist are fully confident in technology (V. Tripodi).

TRUST: Trust or trustworthiness is part of credibility. One suggestion in the workshop was to handle fact checking and trust locally, and then scale up. This approach appears to yield better results. Automated ranking is not yet fully ripe, but to rank humans on trust and knowledge vectors is currently being developed (S. Shulman).

RELEVANCE: This might be addressed in terms of topic, geography, or time, etc. AI helps improve search engines for relevance that is closely related to accuracy: same time, same location, matching results, present factors all pointing towards the veracity of the sources.

DIVERSITY: Diversity not only concerns sources but also plays an important role concerning channels and languages. News information can come from text, image, video, written or spoken language, and in different languages. This variety suggests the need for powerful content management devices that can transcribe and translate on the fly. For gisting, this is already possible but for accurate quotations, full automation is not yet there.

LI Support for Newsroom Activities

a) Information Collection

A journalist collects information to write a story by reading existing content, talking to actors in the field, recording notes, and cross-checking successive drafts with other information sources such as archive material. This information can be in audio, video and textual media.

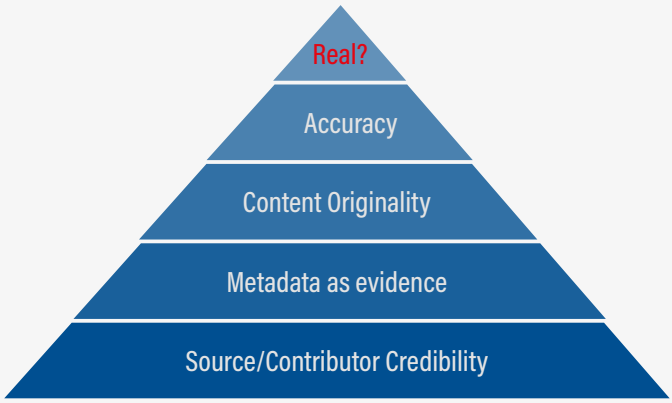


Figure 3 Information collection pyramid: from discovery to verification

- **Text & Voice Search & Fact Checking**
  - Searching online for news and facts about news entities
  - Searching other journalists' stories and sources to understand news stories
  - Searching voice streams for all of the above (voice search)
  - Identifying entities & actions in large text corpora
  - Fact-checking to counter disinformation

- **Voice/Speech Transcription**
  - Transcribing spoken stories from live-streams, recordings, podcasts or radio into written form for ease of access
  - Transcription will join forces with smarter LI to identify relevant topics during a conversation and check information about them online in near real time.

Using LI, search engines inside news work environments may be empowered to handle more functionality, such as making further recommendations to responses to search terms, based on *semantic similarity* or space-time similarity of an event, or other form of relevance.

A radio and/or TV broadcaster will also want to explore the utility of *spoken* search for a large database of recordings (inputting a name or a key term and having the system return audio files that contain such a term along with the time code).

The best way to access information in audio is to have it *transcribed automatically* into written format. There are currently tools on the market that can perform transcription effectively with a very low error rate.

In the future, we can expect the arrival of more semantically-aware search engines, providing insight into content on the basis of conceptual features that can be understood by the engine.
These additional powers of search will largely enable journalists wishing to check on instances of supposed fake news to rapidly compare the disinformative element against the presumed truthful accounts of the same items.

Example of AI/LI automated process for information collection:

Associated Press is currently investigating an automated verification process for UGC (user generated content), AP Verify. It includes: Discovery (search in social media, other news media, etc.), identify source (assess the credibility of the source, find the earliest version), verify content (location, date, time, consistency with other sources etc.) and contacting source (for potential licensing).

A further extension of the power of search will be to combine audio, video and textual search so that the system can return items of spoken interviews and sequences of video, as well as enhanced text search over "fuzzy" characteristics.

It should also be understood that as AI enters the infrastructure of business IT, and penetrates the specific work environment of journalism content management, search terms and results will be increasingly used as *data feeds for further machine learning*. This will in due course enable the effective profiling of individual search records, and thus begin to predict the kind of information that an individual journalist tends to seek (personalized search). Such models will keep track of different information needs over time and react flexibly to changing

information requirements along the way. In this way *proactive* search, able to link together active "words" or concepts in an ongoing investigation, will offer a formidable aid to information management in many cases.

Fact-checking challenge, Technical University Darmstadt

Automated fact-checking presents a set of challenges to Natural Language Processing (NLP), such as document retrieval, stance detection, evidence extraction, and claim validation. These tasks are to be tackled across different domains and heterogeneous textual sources, such as news, debate forums and social media, where most of the controversial claims emerge. Current academic NLP research is aimed at creating a unique corpus resource for training machine learning-based models, based on the fact-checking website Snopes. The first experimental results for modelling fact-checking computationally, based both on the Snopes corpus and the recent Fact Extraction and VERification (FEVER) shared task (http://fever.ai) yielded positive results.

The multilingual layer

Due to the increasingly global scale of information sources, events and existing content, a machine translation (MT) tool should always be available to broaden and objectivize the sources. However, as V. Tripodi pointed out, the most promising approach is to take fact checking and trust local, and then scale up.

Recent developments in machine learning-based translation systems (known as *neural machine translation*) have considerably increased the quality of MT, and contributed to a new dynamic in general for accessing and handling multilingual content sources.

Translation may be used as a *gisting* tool for search - finding out whether, for example, a given text source is relevant by obtaining a rough translation of the main nouns and verbs. For this, existing online solutions are often perfectly adequate, especially as they cover a broad range of major content languages. However, if publishable quotations are to be taken from such content, it will be necessary to double check the accuracy of the translation. In addition to pure language skills, the cultural context likewise needs to be accounted for. Speaking the language does not automatically mean that one 'speaks the culture'. Cultural context may provide essential meaning going well beyond the surface of the word itself.

A critical feature of the MT landscape today is that constant quality monitoring is leading to constantly improving translations over time (by recycling and learning from good translations). The more a journalistic-type discourse (i.e. containing proper names, quotations, event sequences, certain verbs, special vocabulary, turns of phrase, abbreviations, etc.) is used by a given MT engine, the better the translation quality will become. However, it is also a fact that different components are at different levels of readiness for different languages (Andy Secker). This must also be addressed when opting for multilingual information sources.

Images collection & fake detection

The rapid development of new technologies and their decreasing costs used for audio and video manipulation will require adaptation of critical skills to deal with quickly changing information environment (N. Komendantova). Thus, it becomes increasingly difficult to detect manipulated images or video sequences.

There are tools to shorten the distance between image detection and validation but no full image recognition tool is available (V. Tripodi). *The goal is to make the portion of elimination quicker and better*.

Digital fingerprints on photos might be part of a solution.

The InVID project[3] provides a tool for journalists to verify photos and videos, the InVID Verification Application. It is a web-based integrated toolset for the verification of newsworthy user-generated videos and their context.

b) Leveraging the News Database

Any data contained in the collective news database can be rendered far more accessible by using word and phrase-based enhancement tools, which add meaningful tags to pieces of language to identify relevance during analysis. These tags help speed up search, and enable richer, deeper comparisons between the significance of data streams. A key use for this is in *sentiment analysis* of social media, estimating reaction percentages from readers and commentators. Here are the most prevalent tools:

Content Enhancers and Entity Extraction: Data can be enhanced automatically (using learning software) to highlight and extract key values and entities such as dates and people, or key industry language in each domain of interest (engineering, financial, legal, hiring, sport, customerX, etc)

Automatic Topic Classification: This automatically classifies items of content in the database into logical or semantic "topic" groupings (such as political appointments vs. corporate appointments, or football vs. rugby transfers) so that journalists can rapidly focus only on documents relevant to the task at hand. Categories of topics may be pre-defined or developed automatically by machine-learning methods.

Advanced Data Visualisations: Journalists can develop a richer understanding of their data by using such techniques as co-occurrence diagrams and word clouds (based on either frequently occurring words and phrases or user-defined search terms).

- **Sentiment Analysis**
  - Evaluating simple positive/negative sentiment spectra in social media streams around news topics
  - Evaluating spoken content reliability from voice quality & psychographics (e.g. using emotional characteristics of spoken content to estimate sentiment, bias, etc)

c) Writing & Publishing the Story

There are various ways in which LI can help journalists be more productive by writing more quickly and accurately, and outputting news content as text and audio more efficiently.

Apart from *Typing Augmentation* that basically operates in the same way as predictive typing on a smartphone, but is tailored to the style and content of the individual writer and work on a broader level than simply word formation, the main support lies in NLG – **Natural Language Generation**.

NLG uses a combination of machine learning and rule-based techniques to correctly formulate statements in writing a linguistic account of typical *numerical data* such as earnings reports, sports results, weather forecasts and similar numbers-driven information. This makes it useful for small news production companies that wish to maintain a constant flow of news on such topics as the economy, sports results, election results, new appointments, weather, stock markets, traffic figures, etc. without overloading their staff. It is very likely that as data collections grow to multi-billions of examples of such formatted news story types, NLG will be able to learn how to generate entire stories on many other data types than the purely numerical or list-like.

[3] https://www.invid-project.eu/

- **Text Generation**
    - Create readable texts automatically on the basis of a numerical data source such as financial data, sports results, etc.
    - Creating texts that describe contents of videos or images
    - Adapting generated texts to reader profiles to personalise information and address needs

Summarisation: Closely related to the technology of generation is the process of automatically summarising verbal (or audio and even visual) content to provide rapid access to longer documents or collections of documents. Pride of place for summary models would go to the "news summary" that traditionally opens a news broadcast on TV or radio.
Summarisation software exists, using either *extractive or abstractive (sometimes also referred to as constructive or generative)* methods. In *extractive* summaries, the machine identifies the key sentences in a document (or collection of related documents) and creates a summary based on extracting those sentences.
In abstractive summarising, a completely new text is generated on the basis of analysing and re-synthesising the entire document.

Voice news for podcasts, video and radio
There is increasing exploration of voice as a news channel related to non-radio online services.
Podcasts can be generated by text-to-speech (TTS) technology from written output to add a further channel to a broadcaster's offering and to multiply productivity on special occasions. TTS software can now be calibrated to imitate a given speaker's natural voice and then used automatically to produce content with that voice from any written source. This can help enhance a journalist's personal reputation as a "voice".
Commentary to video news coverage can also use the same technique, outputting a journalist's voice based on written news commentary.

Translating news and subtitles: Video news streams will need to use *captioning* (subtitles) to:

a) make the spoken content available to read-only situations, and

b) enable an easy translation of the commentary into other languages.

Subtitles and captions can be translated automatically (and the build-up of language data over time will improve any such service). The written subtitle/caption can also more easily translated into another language.

- **Captioning**
    - When publishing spoken news stories online, either to be audience-inclusive or for foreign language understanding.
    - Translate captions to attract foreign readership of videos.

d) The Virtuous News Data Cycle

If news data lies dormant in a database once it has hit the headlines, then only half of its value is delivered. Currently, most news organisations with an online presence provide links back to earlier stories for background and a richer context. This can add some value to the news service in general.

However, in a digital context, data can take on a further value, especially when combined with the power of LI. There are therefore many reasons why news content should be transformed into more lasting news data that can then be processed in various ways to form intelligent input to a broader range of productive target uses:

1. Automating some types of event *prediction*
2. Constantly providing existing facts over the journalist's shoulder by automatically feeding relevant information on a specific story
3. Constantly identifying factual and other anomalies in online feeds that might signal the work of disinformation
4. Training new journalists by automating the "news journey" as a storyline
5. Creating new services for such future applications as virtual reality enrichment and specialised, interactive news services via home voice hubs for different members of the family.
6. Constant monitoring could help correct/inform colleagues and provide clues in a much more immediate fashion and thus also have a positive impact on collaboration.

To achieve this, database content will need to be enriched, with tags that identify strings of words and phrases that make sense to the machine learning tools (in a sense of allowing these to be used as features for ML). In other words, they will need a comprehensive *semantic framework* that will provide meaningful data choices to any application that automatically consults the database.

In time, therefore, these enriched news databases will be able to help journalists and others predict to a certain degree potential political, medical or social breakthroughs, successes, conflicts and hotspots, or other events of special interest, based on the accumulation of information and the application of inference rules. Journalists will then be able to program their database by setting subject-matter alerts in a far more sophisticated way than is possible with simple word-based search results today. In this way, journalists could be automatically alerted as to (re)emerging stories or potential leads and play a much more informed role in the advance of informed social and political progress.

e) Challenges and potential adverse attitude towards LI in newsroom activities

The News Media sector is traditionally not an "early adopter" of new technologies. Recent research (V. Ertelthaler, University of Vienna) showed why journalists, editors and publishers are not immediately jumping on the opportunities of Language Technologies and AI. The following risks were identified:

- Lack of transparency (function and power)
- Lack of data-integrity - false reports
- Manipulation of datasets
- Self-optimised and closed feedback cycles
- Missing use-knowledge of algorithmic elements

Fake news is not an issue, compared to sheer survival.
Ch. Rainer, PROFIL

However, there is a misconception about AI linked to the manipulation of data sets. If a journalist ended up using manipulated data sets, s/he would involuntarily and unknowingly create "disinformation".
To reduce the risks, human control is needed to guarantee transparency. Technology should only be adopted when manipulation risks are reduced. But there is also a genuine human fear - the loss of jobs due to AI. This is, however, not happening yet. Currently, AI technology is more likely to support journalism, in particular for fact-finding and checking. However, data quality in journalistic AI applications is a serious concern.

In the end, journalists will need to become "technology literate" and this will tend to create new types of jobs rather than destroy them. What is needed is not robo-journalism but AI-empowered journalism whereby AI should be read as "augmented intelligence" rather than the usual "artificial intelligence".

LI Support for News Media Business Models

The market for traditional News Media is shrinking dramatically. The circulation of daily newspapers in Germany is decreasing every year, and advertising in magazines declines 10% a year!
While robo-journalism may drive down costs, it does not help real journalism. Therefore, it is necessary to find other avenues, e.g. people in the news rooms who use the technology to support serious journalism. We need professionals how know to use the technology in the right way.
Investigative journalism is very expensive. Tools for data journalism could be partially used to support it.
What News Media need is LI-enhanced journalism, not robo-journalism.

a) Distribution Channels

News distribution and delivery is not up to speed with other media. It would be great to have something like iTunes or Spotify for news. With sources from different countries and content translated on the fly.
However, there may not be a mass demand for this.

One initiative in the area of new distribution methods is Blendle[4] , a Dutch startup that tries to provide news-articles-on-demand from several Dutch, German and American newspapers. However, it has not (yet?) reached a critical mass and its website is still in Dutch only (with no easily-findable English option).

b) Personalised News Content

A final challenge is the global trend towards the greater personalisation of information provision, inevitably enabled by the proliferation of big data and machine learning, which enables targeted news items. News stories could be delivered as packages that can adapt to facts about the reader (gender, age, interests, delivery system, optimum time for reading/watching, etc) in order to tailor the contents to unique purchasing profiles.

The danger of this trend is that personalisation will enable a news delivery system that could fracture the current problem of fake news into multiple types of disinformation As a result, deciding on what is true or false could well grow even more complex, because different individuals will receive only a personally-biased version of the news in the *language and style they prefer*, making it harder for them to make a rational decision about the value of a given piece of information. My news will not be your news. Readers will be locked into their personal news-bubble. Locating the truthful centre in debate will become even harder.

c) Technology & Media Literacy

Major emphasis is put on the importance of education in terms of media literacy – from the consumer side – and technology literacy – from the journalist side.

Education of the public at large is another way of building resilience among citizens, end-users and voters.
The purpose is to bolster the prevention and mitigation of fake news and reduce the appeal of disinformation and conspiracy theories. Media literacy can build on media and information know-how to foster critical thinking about propaganda and advertising, backed by safer Internet practices for responsible social media use [5].

[4] https://blendle.com

[5] Final report of the High Level Expert Group on Fake News and Online Disinformation, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

However, the psychological element of disinformation and its attraction to user groups (see above 2.) will require a wide-spread initiative across borders and for all age groups.

Regarding technology literacy for journalists, the most important barrier is a psychological one:

a. Get to know the tools and therefore, the risk (or not) of manipulation of data. LI tools must be easy and intuitive to handle and explain, and journalists must "buy in" and appropriate them for their benefit.

b. Journalists need to realise that AI/LT/LI does NOT take away jobs (job loss is due to loss of income by the publisher which is due to loss of subscriptions which is due to the absence of new business models!) but shifts them: more efficient discovery → (part) automated fact check → more time for real investigation. Hence the need for "Data-journalism" not "Robo-journalism". Hence, targeted LI tools integration into existing journalistic content management tools AND training is needed to benefit the most from technology.

LI for the News Media – Challenges for the near future

The graph shows the News value chain and its corresponding needs. As can be seen, all parts in red are not yet (fully) met by technologies.

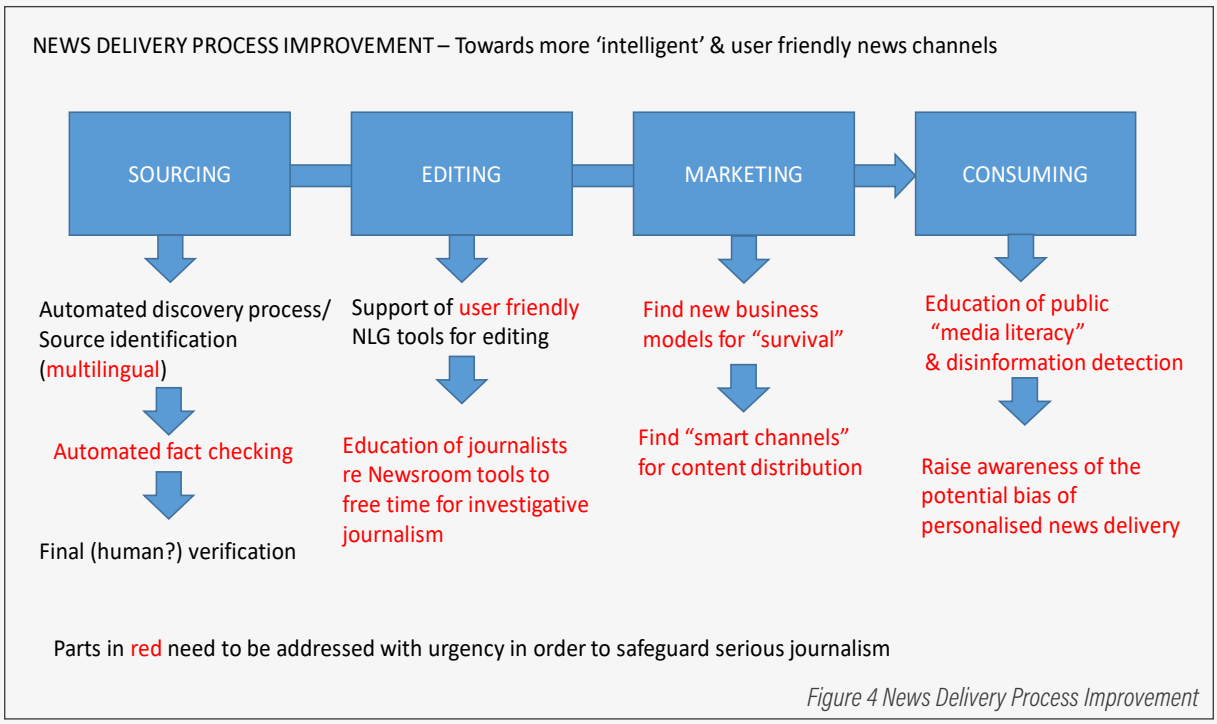


Figure 4 News Delivery Process Improvement

Challenge 1: Multilingual, cross-channel discovery and source identification. It might sound trivial but it is not: Speech recognition, transcriptions, translations, contextualization in multiple languages is not an easy job. While machine learning is increasing its capacity, and contextualization the huge set of data needed is still an impediment. In addition, applying ML on the wrong data may lead to biasing and counter-productive effects.

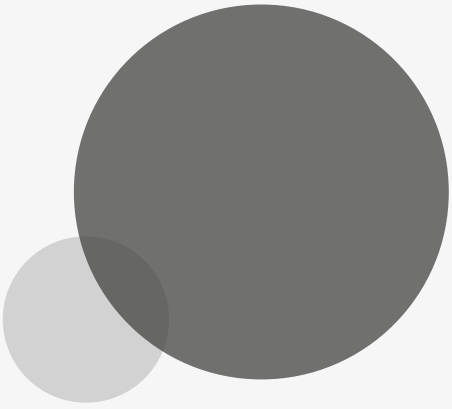
A similar problem affects

Challenge 2: Automated fact checking: Only comparable texts or visual comparisons can guarantee accuracy. Some say that a final human verification will always be needed. Human-machine interaction, i.e. a combination of automatic processing and human-interpretation, balanced in an adequate manner, may provide a promising approach. The integration of discovery and fact checking tools into existing content management tools for journalism is yet another challenge, related to both, 1 and 2.

Challenge 3: NLG can make the life of journalists easier but this tool needs to be more user friendly (a word pretty alien to techies, even for simple solutions). Apart from that, journalists must be trained to use such tools and be convinced that it helps rather than impedes their work (see above, c).

Challenge 4: Traditional news media are fighting for survival. Online versions are free of charge most of the time, and paper subscriptions are going down the drain, together with advertising income. New business models could be automated "read-on-demand" or personalized content. However, a Spotify for News remains to be created.

Challenge 5: Educate the public about media literacy, disinformation and the potential bias personalised news might create. Only enlightened users can support the combat against disinformation effectively. There are already legal and technology measures available in many countries designed to root out "misinformation" and ultimately "disinformation" in online journalism and news broadcasting, especially at State level. Some of these efforts also address other issues such as propaganda[6] . Clear guidelines on the nature and methods of disinformation i.e. "intentional bias in news content production" should be spread on a wider basis, so that younger generations and communities at risk can arm themselves against such a threat.



[6] See, for example, https://www.poynter.org/ifcn/anti-misinformation-actions/

Annex

1 Content & Method

This Orientation Paper is based on the findings of the conference "Fake News and other AI Challenges for the News Media in the 21st Century" that was held on 29/30 November 2018 in Vienna, Austria. Furthermore, it uses background information from related articles and reports (Source: https://www.scoop.it/topic/language-tech-market-news/?&tag=news+media)
The concept can be visualised as such:

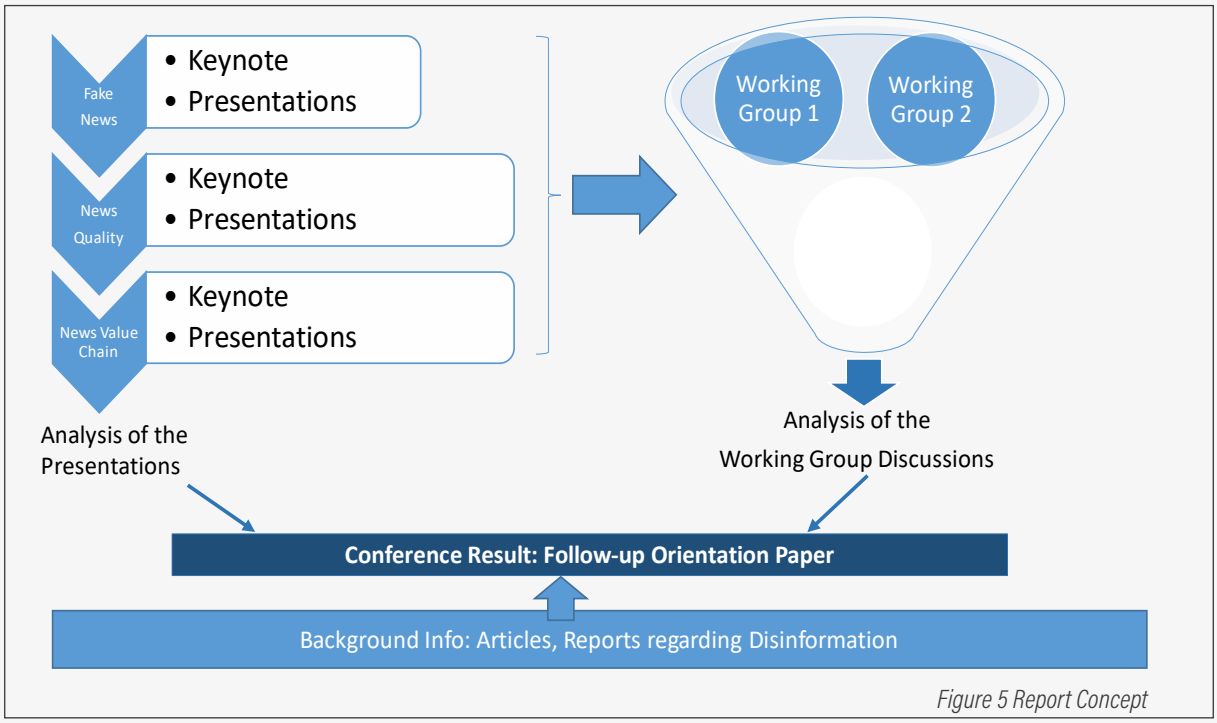


Figure 5 Report Concept

2 Programme of the "Fake News..." conference & other sources

Event website
https://www.liatwork.com/fake-news-and-other-ai-challenges-for-the-news-media

Programme and presentations:
http://www.lt-innovate.org/sites/default/files/DOWNLOADS_0.pdf

https://www.sail-labs.com/2019/03/01/the-summary-report-i/

https://www.sail-labs.com/2018/12/06/fake-news-and-other-ai-challenges-for-the-news-media-in-the-21st-century-conference/

Other Sources:

Final report of the High Level Expert Group on Fake News and Online Disinformation
http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

LangTechNews articles on Language Intelligence and the News Media
https://www.scoop.it/topic/language-tech-market-news/?&tag=news+media

LangTechNews articles on LI and fake news
https://www.scoop.it/topic/language-tech-market-news/?q=fake+news

Data Society report on Media Manipulation
https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

3 Current (selected) Initiatives to combat Disinformation

The InVID project (Horizon 2020) provides a tool for journalist to verify photos and videos.
https://www.invid-project.eu/

The SUMMA project (Scalable Understanding for Multilingual Media) enables multilingual media monitoring at scale.
http://summa-project.eu/

The FANDANGO (Fake News discovery & propagation from Big Data Analysis and AI operations) project
https://fandango-project.eu

https://www.snopes.com/about-snopes/ Fact-checking website (US) used by the University of Darmstadt
http://fever.ai Conference on how to do fact-checking

Digital News Initiative by Google in Europe
https://newsinitiative.withgoogle.com/dnifund/
EC Code of Practice on Disinformation
https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

Action Plan against Disinformation – EC (also in view of the upcoming European elections)
https://ec.europa.eu/information_society/newsroom/image/document/2018-49/action_plan_against_disinformation_26A2EA85-DE63-03C0-25A096932DAB1F95_55952.pdf

EU vs. Disinformation campaign (set up by Member States in 2015)
https://eeas.europa.eu/headquarters/headquarters-homepage/2116/-questions-and-answers-about-the-east-stratcom-task-force_en
https://euvsdisinfo.eu

THANKS TO OUR SPONSORS

The conference «Fake News and other AI challenges» would not have been possible without the support of the following sponsors:

- Vienna Business Agency
- European Open Source Intelligence Forum
- European Marketing & Financing Services
- 24translate

The report was edited by Andrew Joscelyne (LT-Innovate), Margaretha Mazura (EMFS), Philippe Wacker (LT-Innovate), Gerhard Backfried (SAIL LABS), Dorothea Thomas-Aniola (SAIL LABS).

Without the presentations of the speakers of the conference and the workshop discussions, this report would not have been possible. The speakers were, in alphabetical order:

Gerhard Backfried, SAIL LABS; Peter Cochrane, University of Suffolk; Gerald Czech, Red Cross Austria; Victoria Ertelthalner, University of Vienna; Markus Glanzer, Red Cross Austria; Iryna Gurevych, Technical University Darmstadt; Allan Hanbury, Technical University Vienna; Andreas Hanselowski, Technical University Darmstadt; Mustafa Isik, Kerngedanke; Nadejda Komendantova, IIASA; Vassilis Kappis, University of Buckingham; Florian Laszlo, FIBEP; Gari Own, EUROSINT; Bettina Paur, University of Vienna; Mark Pfeiffer, SAIL LABS; Christian Rainer, PROFIL; Vladimir Sazonov, University of Tartu; Andrew Secker, BBC News Lab; Stuart W. Shulman, Texifter; Vincent Tripodi, Associated Press; Kristof Varga, Bakamo Social Public; Andreas Ventsel, University of Tartu; Philippe Wacker, LT-Innovate; Rosemary Wolfe, Industry Expert.

The event was co-organised by:

SAIL LABS and LT-Innovate