ZA Cloud SRL, str. Govora 16A,
400664 Cluj-Napoca, România,
Tel: +40 723 175 635,
Email: office@zettacloud.ro

White Paper

Automated Approaches to Digital Content Verification

***We live in the age of mass online content consumption, reaching the status "information obesity" epidemic foreseen by author Avlin Toffler**. As a content creator or distributor, it has never been easier to create and propagate information, true or false. However, for the content consumer, it has never been more difficult to verify the trustworthiness of online content and make an informed decision.*

What are the different types of untrusted content?

From news websites, blogs, message boards and social media platforms, untrusted content ranges from the seemingly innocent **satire** to government sponsored **disinformation propaganda**. In between, we deal every day with **clickbait** content, **source hacking**, **partisan** news, **sponsored** content, **conspiracy theories** or plain **journalistic errors**.

The current approaches to fighting #fakenews

Professional Fact Checkers: Verification performed by media professionals, with accurate results following a standardized approach and methodology. Time consuming, hard to scale, little impact.

Crowdsourced Fact Checking: Verification performed by non-professional communities, leveraging "crowd wisdom" and with potential to scale. Easy to troll, detour and become partisan.

Non-assisted Automated Verifications: Software-based content verification based on statistical algorithms mimicking the critical thinking process. Huge scaling and impact potential.

The Artificial Intelligence Advantage

Natural Language Processing (NLP), a discipline that combines Computer Science and Artificial Intelligence, has been proven to effectively help humans interact with online content. From search engines to bots, NLP has evolved into a ubiquitous tool that can analyze digital text as accurately as a human being would.

We rely every day on AI algorithms to search for content, get product recommendations, keep our email free of spam or even get medical diagnosis. Using the advantages of AI for flagging untrustworthy content is the obvious next step.

The TrustServista Approach

In 2017 Zetta Cloud launched the first ever platform to automatically verify the trustworthiness of online news:

"Uncovering the hidden part of the information iceberg"

www.trustservista.com

Embedding more than a dozen NLP-based algorithms, TrustServista has been proven to effectively flag untrusted content in real-time, helping analysts shorten investigation times with a 3-step verification approach:

Publisher Veracity

Each analyzed webpage contains valuable meta-data that impacts the veracity of the publisher: a named author, a verifiable digital "fingerprint" (public DNS ownership, unique advertising code snippets) and contact information. **In time, with enough content collected from publishers, a statistical publisher "persona" can be defined, answering the question of "Who is publishing this?"**

Information Source

Every article published online has one or multiple sources of information: news agency wires, other websites, social media or journalists- recorded events or statements. Automatically building a relationship graph from each article - containing the links to other webpages, references and close similarities with other articles - can be done using a combination of NLP, Graph Theory and

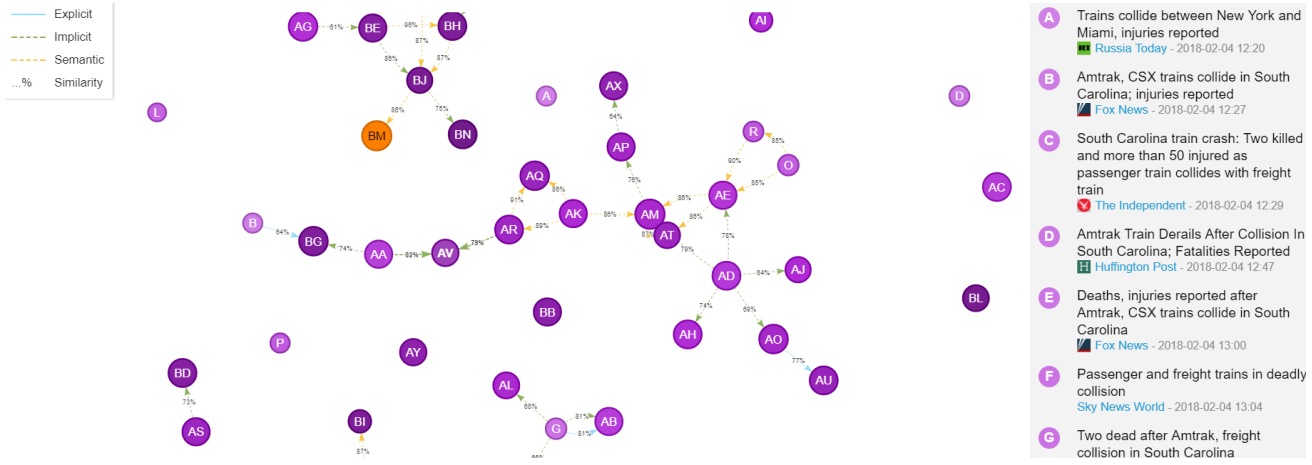
Semantic Comparison. **This way, all the information sources of an article can be traced back to "Patient Zero", the potential origin of information answering the complicated question of "Where does this come from?"**

Content Quality

The 3rd and most important verification is done on the content itself, as writing style is directly correlated with trustworthiness. Making heavy use of NLP algorithms, the **polarity** of the article can be determined (positive, negative, neutral), measuring the author's subjectivity; the amount of factual information can be extracted as a **context setting** (Who? When? Where?) formed by named entities. A unique statistical algorithm can determine if the article is written in a **clickbaiting** style and the content is automatically categorized according to the **IPTC** and **IAB taxonomies**. **This way, the content analysis can reveal the answer to the question "How was it written?"**

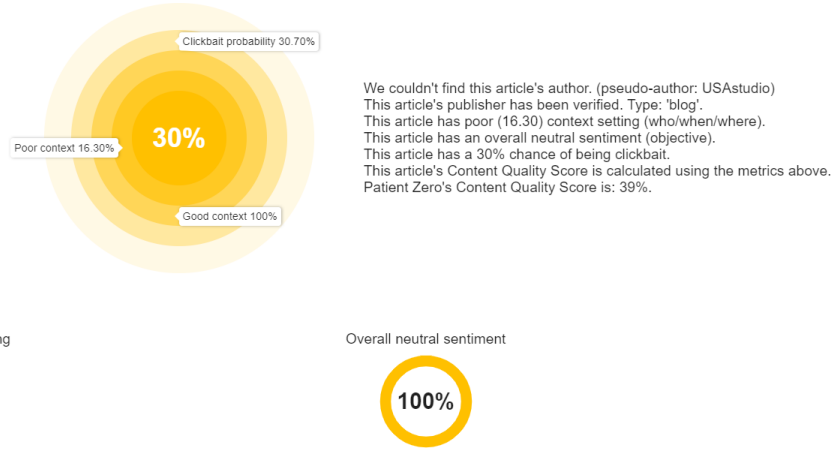
An overall trustworthiness score can be calculated based on the Publisher Veracity, Information Source and Content Quality metrics, automatically and with no human assistance, relying entirely on AI software algorithms.

In addition, **Context Information** can further define the overall trustworthiness score, by analyzing similar content, using Semantic Similarity algorithms applied to large sets of content data:



Trustworthiness score explained

The current formula for determining the trustworthiness score relies on the 3 types of scores detailed previously: publisher, source and content quality.



A Trustworthiness score can generally be defined using the following base formula:

Intermediate Score = (Publisher Score * Publisher weight) + (Content Quality Score * Content Quality Weight)

The **Final Score** can also take into account the **Source Score**, to **increase** (if the source exists and has a high Intermediate Score, or if the analyzed article is actually the source) or **decrease** (if the source does not exist or it exists and its Intermediate Score is low).

The weights applied to any of the scores are variable and can be determined based on experiments. For example, if the Content Quality Score is most important for trustworthiness calculation, it will be given a higher calculation weight. The elements that allow the definition and calculation of the scores are:

Publisher Score

- The **type of the publishers**: a recognized media agency or publisher versus a blog or an anonymous publishing platform.

- The publisher's identifiable legal status and contact details: if a legal entity name can be determined for the publisher, with clear indication of the website ownership, together with contact details such as address and phone number, versus generic web-based contact forms or ownership that cannot be determined.
- The publisher's **web hosting location**, which is usually either a global web-hosting provider or a local one, located in the same country/region as the publisher's declared location or audience targeted location, versus hosting of websites in different countries.
- The **authors of the content**, identified for each article, should at least be a named author (first and last name), with contact details and traceable social media profiles, versus generic or missing author names.

Content Quality Score

- **Clickbait score**: calculated by analyzing the writing style of the title (use of heavy punctuation, uppercase letters), its sentiment and lack of named entities that provide meaning and information value to the title itself. The information from the title is then compared to the information found in the article body, to determine if the title is not misleading and matches what is found in the body of the article.
- **Content setting**: a calculation of the number of the relevant named entities found in the article body (names of persons, organizations, locations, titles and so on) versus the article body length. If the article has a good (above 60%) representation of named entities in the body, then it is considered that the content has context, meaning it contains enough factual information. If named entities are not found in the article body, or the number of occurrences are too low compared to the article length, it is considered that the "context" is not well represented and the article does not contain enough factual information.
- **Sentiment**: the determination of the articles polarity - positive, neutral, negative - determines either the "sentiment" of the topic of the article (war, conflict) or the sentiment of the author. This metric should be also compared with the automated classification of the article, or with the author's stance (a metric not yet developed in TrustServista), to provide a clear and meaningful indication of subjectivity vs objectivity of the author as represented in the sentiment analysis.

Source Score

The source score relies on tracking the information source back to its Patient Zero. There are a number of scenarios to be taken in account:

1. **The article does not have a Patient Zero**. In this case, the final trustworthiness score can either take this in account and be lowered, or do not include it altogether.
2. **The article is Patient Zero (for other articles)**, and the final score does not take this into account.
3. **The article's Patient Zero has a low intermediate score**, which also lowers the final score.
4. **The article's Patient Zero has a high intermediate score**, which also increases the final score.

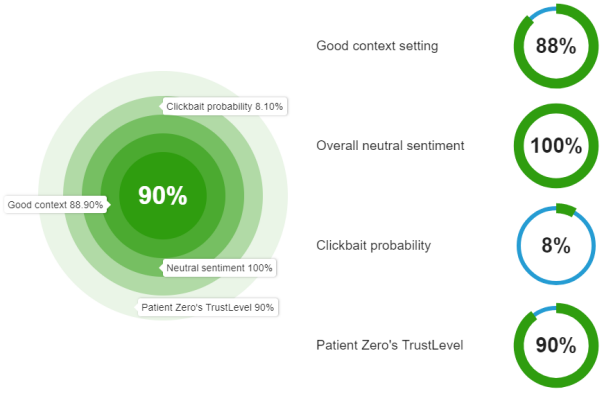
Other metrics that can be calculated and added the the trustworthiness score are:

- Article image analysis (if the images have been published elsewhere on the web or have been altered) or deepfake video analysis.
- Author stance detection.
- Historical publisher/author analysis for the article domain and key topics.

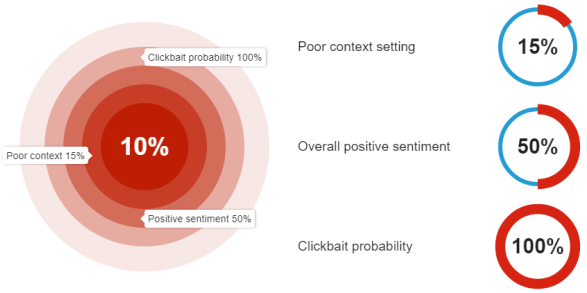
Read here the TrustServista news verification report:
https://www.trustservista.com/2017/08/trustservista-publishes-its-first-news-verification-report/

Example

An article that is well written, contains factual information, is written in a neutral way, does not make use of clickbait content and has all the credentials available will receive a high trustworthiness score. On the opposite, unnamed authors, clickbaiting titles and no context will lead to low trustworthiness score.



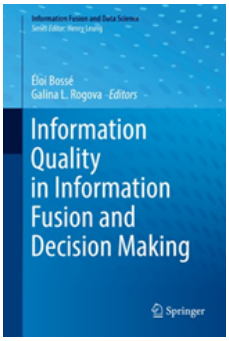
We found this article's author (Ben Westcott, Steven Jiang and Serenitie Wang, CNN).
This article's publisher has been verified. Type: 'newspaper'.
This article has good (88.90%) context setting (who/when/where).
This article has an overall neutral sentiment (objective).
This article has a 8% chance of being clickbait.
This article's Content Quality Score is calculated using the metrics above.
Patient Zero's Content Quality Score is: 90%.



We couldn't find this article's author. (pseudo-author: jawe)
This article's publisher has been verified. Type: 'blog'.
This article has poor (15.00%) context setting (who/when/where).
This article has an overall Positive sentiment (subjective).
This article has a 100% chance of being clickbait.
This article's Content Quality Score is calculated using the metrics above.

Read more Case Studies here: https://www.trustservista.com/category/case-studies/

Further Reading

	**"Fake or Fact? Theoretical and Practical Aspects of Fake News"** **Authors**: George Bara (Zetta Cloud), Gerhard Backfried, Dorothea Thomas-Aniola (Hensoldt Analytics) **Date**: 2019 **Publisher**: Springer International Publishing **Link**: https://www.springerprofessional.de/en/fake-or-fact-theoretical-and-practical-aspects-of-fake-news/16600900
IEEE CogSIMA™	**"Building a Dynamic Corpus of Fake News Using Commercially Available Machine Translation and NLP Software"** **Authors**: George Bara **Date**: 2021 **Link:** https://edas.info/showManuscript.php?m=1570703796&ext=pdf&type=stamped

Conclusion

The Internet is full of fake news stories. The tremendous volume of news items created every day in multiple languages and distributed across many channels makes schema-based fact checking useless. There is no efficient way that any of these approaches can solve the issue of determining the trust of new stories developing in real-time, with impact on how fake news or propaganda spreads and influences readers.

There is a critical need for determining the trustworthiness of news stories automatically, through specialized algorithms. Concepts of critical reading and investigative journalism can be automated; mass data can be collected, structured, categorized and labeled. Articles can be linked between each other to determine where information originated. Text Analytics algorithms can extract named entities, determine similarities, label the sentiment and put every single news article into perspective, understanding if it has those key elements that will brand it as being "trustworthy" or not.

Technical Specifications

Delivery Models

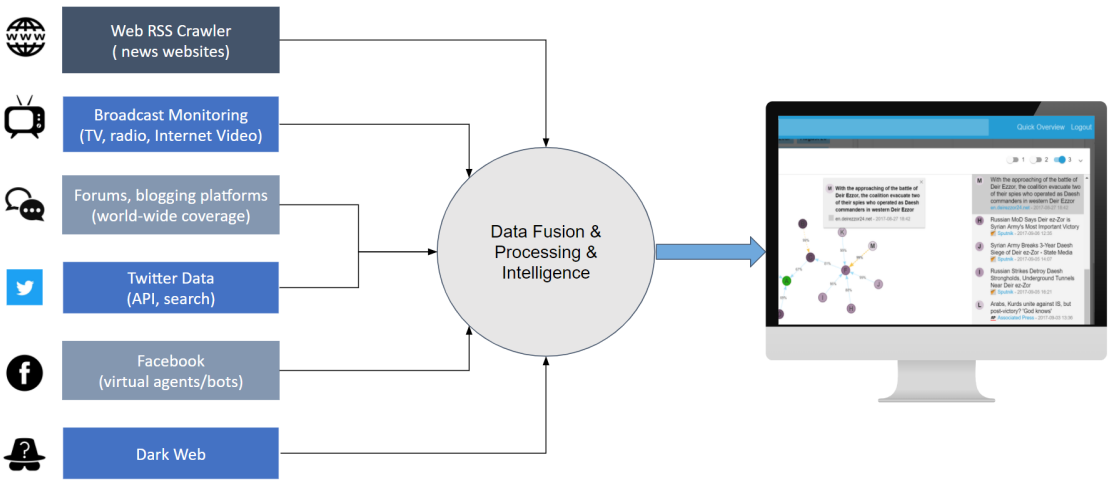
TrustServista has a very flexible delivery model that addresses the needs of every customer persona:

- **Multi Tenant SaaS**: no installation required, just log into our platform with a username and password.
- **Private Cloud**: we can deploy a highly scalable instance of TrustServista on your choice of cloud platform, only for you to access.
- **On-premise**: the most secure option, with the entire TrustServista platform installed in your datacenter.

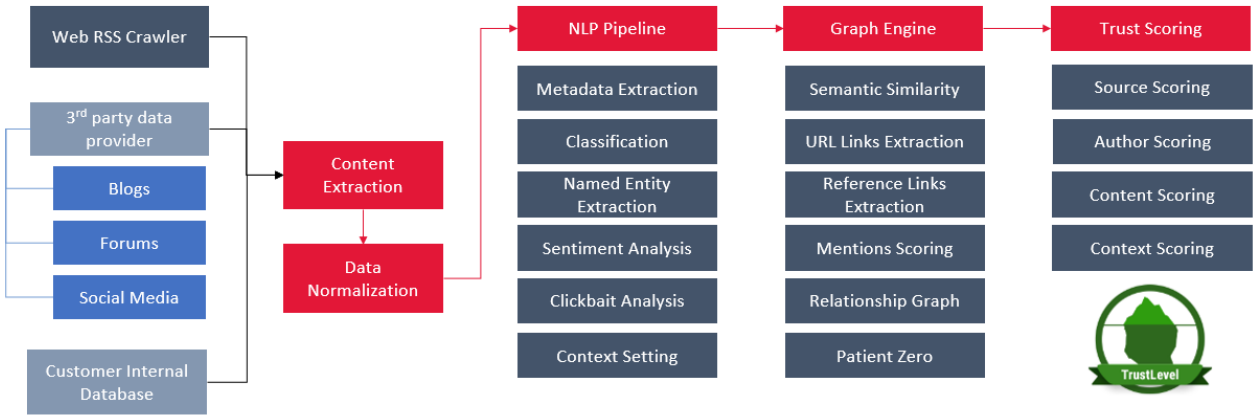
TrustServista can be utilized through:

- A Web Dashboard
- REST APIs
- Chrome extension

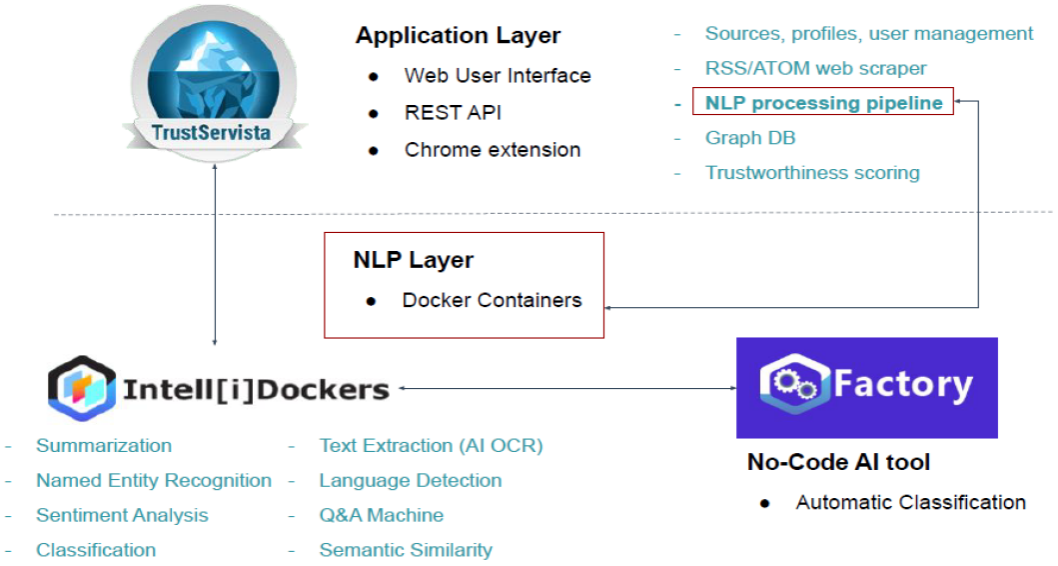
Solution Architecture



TrustServista has a modular architecture, with Dockerized components that perform specific tasks: Data Collection, Data Processing and Analytics & Reporting.



The Data Processing layer is powered by Zetta Cloud's leading NLP solution stack **IntelliDockers** and its Adaptation No-Code AI platform **Factory**:



The core NLP capabilities of TrustServista are provided by the highly scalable and robust **IntelliDockers** technology suite. More information: https://www.intellidockers.com/

All the NLP capabilities can be adapted to specific domains or use cases directly by the customer, with little or no Data Science skills using **Factory**. More information: https://factory-nocode.ai/

Supported Languages

TrustServista has native support for a broad range of languages, with complementary Neural Machine Translation capabilities for unsupported languages.

Native supported languages: Arabic, English, Farsi, German, Hungarian, Italian, Polish, Romanian, Russian.

Language supported for the additional Clickbait metric: Romanian, English, German.

Languages supported with Machine Translation: Albanian, Dari, Hindi, Maltese , Somali Arabic, Dutch, Hungarian, Norwegian, Spanish Armenian, English, Indonesian, Pashto, Swahili, Bengali, Estonian, Italian, Persian, Swedish, Bulgarian, Finnish, Japanese, Polish , Thai, Burmese, French , Javanese , Portuguese, Turkish, Catalan, French (Canada), Khmer, Portuguese (Brazilian) , Ukrainian, Chinese (Simplified), Georgian, Korean, Romanian, Urdu, Chinese (Traditional), German, Kurdish, Russian, Uzbek, Croatian, Greek, Latvian, Serbian, Vietnamese, Czech, Hausa, Lithuanian, Slovakian, Danish, Hebrew , Malay, Slovenian.

www.trustservista.com

