Making SNS and other Digital Spaces more Safe and Sound

Takahiro Okumura
Manager
Marketing Communication Group, Communication Department
Eltes Co., Ltd.

1. Changes and Issues in "Human×Digital Space" with COVID-19

The COVID-19 pandemic has resulted in acceleration of digital transformation (DX). In business, non-face-to-face communication has increased, conducting meetings and customer interaction on-line, where physical distance is not an issue. Communication is moving from real to digital space. Use of online tools is considerable, both for B2B and B2C activities, and the importance of reputation in digital space is expected to increase even more in the future. As such, an understanding of how tone governs digital space is essential for decision making in enterprise management.

However, it is extremely difficult to understand tone accurately in digital space. It is possible to understand and verify tone to some extent by just gathering information with a search engine, but on SNS, where information can flow forcefully and energetically, it is difficult to grasp the tone correctly. This is because distortions of a speaker's intentions can accumulate, and after non-factual information has spread, it cannot be suppressed. This results in a particular risk of damage due to flare-ups or rumors on the Internet.

It is still fresh in our minds how there has also been an "infodemic" with the spread of the COVID-19 pandemic, causing fear and uncertainty for many people. For example, in Japan there was widespread speculation on SNS, that materials for masks would be diverted from Japan and import of materials from China would stop, which resulted in a shortage of toilet paper. In Iran, there was a rumor on SNS that drinking strong alcohol would kill the virus in one's body. Many people who believed the rumor drank strong alcohol containing methanol and died of methanol poisoning. To raise awareness regarding this infodemic, the World Health Organization (WHO) created new content on their COVID-19 Web site, with advice regarding superstition and anxiety and how to avoid being misled by unreliable information.

There have also been cases of widespread rumors, where posts were made with parts of a paper or someone else's post were quoted, cutting out and rearranged phrases to suit the biases of the poster. For example, information stating that "epigallocatechin gallate (EGCG), found in green tea, has strong anti-viral effect," degenerated into "is an effective treatment for novel coronavirus," and got distributed. In this case, the essence of the information was lost by the extraction and rearranging of wording.

Communication on SNS occurs through the exchange of short messages, so brevity results in strong tendencies not to convey the intended meaning, for the meaning to change when only part of a sequence of tweets is quoted, and for information to become distorted in the process of transmission. Readers also skim through the huge flow of messages over time and can interpret messages incorrectly, without carefully scrutinizing the information. People today are exposed to huge amounts of information and are used to consuming information in summary, so it is not uncommon to just accept whatever information is received. Eltes is conducting analyses of information related to COVID-19. In this process we have learned that incorrect information that gets distributed on SNS often dies down due to fact-checked articles appearing in the media, but then it sometimes begins to spread again on bulletin boards and other media. In our diversifying digital space, a single fact check alone will not correct incorrect information, and there is a possibility that it will flare up again. Our analysis has revealed the problem that once incorrect information has spread it transforms successively in digital space and variations tend to increase.

The likelihood that information will be misinterpreted due to poor writing ability or poor reading comprehension is always high, so incorrect information is often repeated, moved and reissued in digital space. For these reasons, when information starts to disperse, it is increasingly difficult to follow in real time.

2. Danger that distortion in digital space will affect the real world

"DIGITAL2020: GLOBAL DIGITAL OVERVIEW" is a report analyzing the trends and tendencies of people around the world on digital, mobile and social media. According to the report, more than 4.5 billion people used the Internet, and users of social media had surpassed 3.8 billion people as of the beginning of 2020. This means nearly 60% of the global population is using the Internet, an increase of 7% compared to 2019. The average Internet user spends six hours and 43 minutes on line every day, so assuming they sleep eight hours a day, 40% or more of their waking hours are spent using the Internet.

The report also indicates that the time connected to digital space is increasing from year to year. The COVID-19 pandemic has spurred this on, and with fewer opportunities to meet face-to-face, we can expect people are more active collecting information online. Although digitization is making people's lives more convenient, it also brings new risks. In fact, as the amount of connection time has increased, the amount of slander and defamation on the Internet has also tended to increase. Cases of arrest or litigation due to malicious posts are becoming more and more common. Cases where thoughtless, slanderous comments made anonymously have resulted in lost lives are also continuing to occur. Calls for morals in digital space are increasing daily.

The idea that digital space and real society is separate is a thing of the past. A person's personality and corresponding reputation in digital space has become inseparable from that in the real world, whether good or bad. Because of this, there may be increasing danger that when distortions appear in digital space, they will have effects in the real world as well.

This is not only a problem for individuals; it also applies to enterprises. Enterprise managers must make certain preparations for the risk of injustices characteristic of digital space. An important factor in doing so is the use of intelligence, the result of information processing and analysis, as the basis of decision making.

When slander, defamation or flare-ups occurs on the Internet, many people have a tendency to perceive it as though it is the prevailing attitude in society. However, such issues cannot be dealt with properly without correctly determining whether it really is a prevailing attitude or just that of a noisy few. Incidentally, during Japan's Warring States period, warring armies were known to use human figures to give the impression that they were stronger militarily. In a similar way, there have been many cases when a single person has created additional accounts on SNS and other platforms to spread slander and defamation, when actually it is just one or a small number of people. Rumors are like natural disasters and impossible to prevent before they occur. However, damage can be minimized by perceiving the matter correctly. To prevent the spread of negative effects in the real world, enterprises need intelligence, including that gained from monitoring digital space.

3. Eltes accomplishments in the domain of Internet flare-up and rumor damage control

Eltes has set a policy of "continuing to fight against digital risk," and as an organization of digital risk management specialists, we are developing various solutions to resolve digital risk. We support management of digital risk emerging as digital transformation (DX) of society progresses, including the expansion of services such as search engines, SNS and on-line banking, made possible by development of information and communications infrastructure technology and digital devices. For example, we provide comprehensive risk management solutions for enterprise social media operations, which have presented real risk as business environments have changed. We have provided digital risk management-related services to over 1,000 companies, including listed companies such as NTT DOCOMO, Mazda, and Suntory.

There are two main types of digital risk management. The first is "Social Risk Management." As the number of SNS users has increased, "incidents" have occurred frequently. The number of Internet flare-ups has increased annually since 2011. We develop support for social risk management to help prevent such incidents in three phases, which are: "Survey, analysis, and system building," in which we gain an understanding of potential risks and business improvements and decide rules; "Operations," in which we perform early detection of emergent risks and rapid initial response; and "Countermeasures," in which we support mitigation of such risks.

The "Survey, analysis, and system building" service involves collection and analysis of information on the Internet regarding the enterprise and its products and services, which we deliver as a report. We also provide a marketing analysis through comparison with competition and information collected regarding leaks, events and incidents occurring overseas. We first collect all articles on the Web related to our client enterprise. These are placed in categories such as positive or negative, according to pre-determined conditions. We also expose potential risks and issues like the detailed reputation of products and analyze future initiatives. These are also summarized in a report. We can also create operating regulations and manuals necessary for introducing a new, public SNS presence, or if social media policies were created several years earlier and are no longer suited to current conditions, we can support revising them and provide other follow-up on the organizational structure of SNS operations.

The main activity during the "Operations" phase is Web risk monitoring. We monitor the Internet 24 hours-a-day and 365 days-a-year for information including rumors regarding the enterprise or its products or services, risk of information leak due to the company's employees, and other particulars such as risk related to malvertising, pharmaceutical products or consumer protection. If urgent information is detected, we also consult regarding urgent notifications and how to deal with them.

For the final phase, "Countermeasures," we provide our Web Risk Monitoring clients with dedicated consultants, who seamlessly conduct a risk assessment for the matter when a risk is detected. The content that precipitated the crisis is analyzed for credibility and potential effect, and a profile of the source is prepared. For escalation after risk is detected, we can even provide support for crisis management public relations as needed, create press releases, train for press interviews and provide advice on handling the situation going forward. If the crisis grows, we also provide consulting services to handle public relations for crisis management. To deal with search engine reputation, we can identify issues with how the enterprise, its products or services, are viewed on search engines, which is an important factor in forming reputation and users' brand experience. We also plan strategies to solve issues and achieve objectives, and design KPIs.

We now discuss some examples of enterprises that have introduced Web Risk Monitoring. Food products company, A Inc., was concerned that they would be associated with a flare-up on SNS regarding the food-products industry in 2016. They decided to use Eltes Web Risk Monitoring based on our "24 hour-a-day, 365 days-a-year risk detection system," and our "ability to provide rapid contact with dedicated staff." That is, the fact that we have been able to implement effective risk monitoring, and that human staff use their intuition in interpreting subtleties in posts regarding the company as they appear each day and in making decisions. Our monitoring service can also support matters other than risk, and not only when negative incidents occur. An example is collecting information regarding reactions after a commercial is broadcast.

Web Risk Monitoring is not limited to monitoring SNS regarding a company's products or promotional activities. Company B in the service industry uses it to detect potentially risky posts on SNS, but also to gather positive feedback from customers regarding service provision, which it uses internally to give recognition and commendations.

In addition to the Social Risk Management service we have been discussing, Eltes also has initiatives for enterprise digital risk counter measures, with two approaches to Internal Risk management. These involve cross-sectional analysis of logs, and analysis of behavior to detect internal behavioral risk.

4. Advanced flame-up/rumor damage control using AI

Eltes is introducing AI to our Web Risk Monitoring Service to improve quality. However, we are not depending entirely on AI. We hope to improve both service quality and efficiency by integrating the strengths of both AI and human operators. As such, we first used AI to implement a mechanism that classifies posts as either negative, neutral, or positive for our Web Risk Monitoring service, as a measure against flare-ups and rumors on the Internet. However, we had three main problems with this effort.

(1) A large amount of correct training data was needed to obtain correct classifications from the AI.

(2) Popular words and phrases come and go, so updates (maintenance) are necessary.

(3) Interpreting the meaning correctly from the text and context is very difficult.

To solve problem (1) we began work creating training data for an AI to classify posts as negative, neutral, or positive, based on the data that we had been collecting continuously since the Web Risk Monitoring Service began in 2011. To create the training data, we had to indicate which posts were obviously negative, or positive, but we found that even for the same post, there are cases when this decision would be different depending on the type of business or industry, so deciding what was "obvious" was not easy. It was relatively easy to collect posts that are "obviously" neutral, but it was much more difficult than we expected to prepare a large number of posts that were negative or positive. Our long experience providing services contributed greatly to creating enough high-quality training data.

For problem (2), that popular words and phrases come and go, we designed a solution involving human intervention. For example, use of the Japanese word, "ataoka" to mean "someone with a screw loose" became generally popular after being used by a comedian, and won 1st prize in the Insta-Buzzword awards announced by Petrel Inc. in the first half of 2019. Until several years ago, "ataoka" was not a recognized word, and that might be why the AI did not classify it as negative. Thus, there are trends in the language used on SNS, and such changes in language must be reflected constantly in the training data used by the AI to classify negative and positive posts. To handle this, we are using human sensitivities to capture the meaning and sense of words, and reflecting this accurately in our training data.

Finally, problem (3), correctly interpreting meaning from text and context, is difficult for AIs. Morphological analysis divides the text into small units to extract the meaning, but even a particle can completely change the meaning, so it is very difficult. For example, the phrase, "I like B better than A," in a post comparing two companies' products tends to be difficult to classify as negative, neural, or positive. Considering that currently AIs also have difficulty with posts that contain positive words but are negative (such as "I always wanted to go to that store, but unfortunately could not"), and posts that contain unusual expressions, we have these posts classified by people.

5. Eltes optimized digital risk countermeasure solution

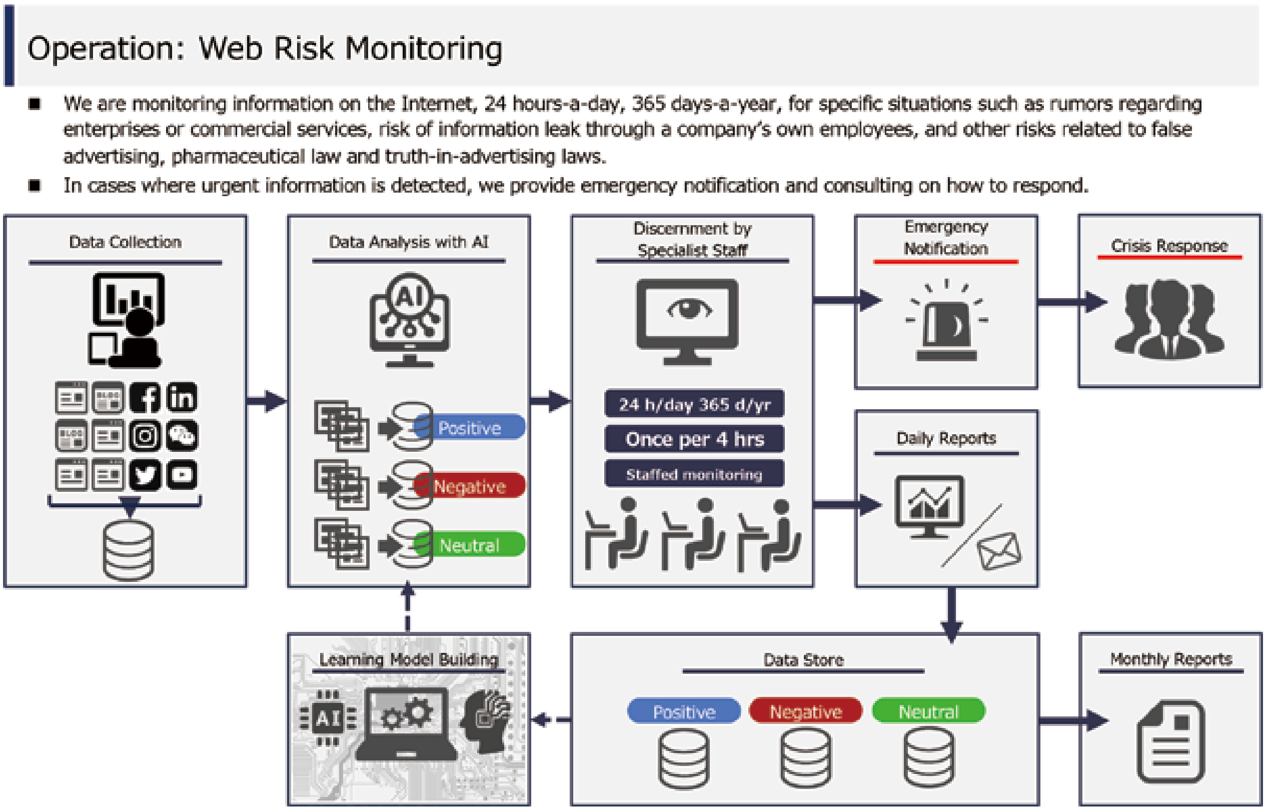
The value of the Web Risk Monitoring Service is early detection of posts that pose a risk. Put another way, negative posts will not go unnoticed. Although 80 to 90% of posts were classified as neutral even before AI was introduced, the problem remains that it is very difficult to interpret the meaning correctly from the text and context, and we found that it is difficult to leave Web Risk Monitoring entirely to the AI. As such, we created a workflow giving the role of screening for neutral posts to the AI, excluding posts that could be negative and are difficult for the AI, such as when the meaning is difficult to interpret from the text and context, and when they contain a mix of positive and negative words. Such posts are handled by a dedicated staff member.

As a result, our human staff concentrate on subtle posts on the Internet that may involve our client and be related to flare-ups or rumors. Thus, we have built an environment in which human error is reduced and we can detect flare-ups and issue urgent notifications quickly, improving the quality of the service. We are also able to use positive and negative data judged by our human staff to train the AI, incorporating information about language trends and improving the accuracy of AI results.

The same post can have widely varying effect on the reputation of an enterprise, depending on the enterprise in question and trends in society. In addition to sending urgent notifications regarding posts that indicate risk to an enterprise, we also offer dedicated consultants that can support the enterprise in its initial response.

It is difficult to judge the effects on real society, of what is superficially just numbers and text in digital space. Distortions emerge in digital space due to the actions of people, and Eltes is fighting against the resulting digital risk. We are using our expertise to combine the strengths of both people and AI, and will continue the fight against digital risk as it continues to change.

■ **Figure: Eltes Co. Ltd. “Web Risk Monitoring Service” Overview.**
24 hour-a-day, 365 days-a-year monitoring of posts for various types of incident, crisis notification and consultation to deal with them when posts including risky content are detected.



Operation: Web Risk Monitoring

- We are monitoring information on the Internet, 24 hours-a-day, 365 days-a-year, for specific situations such as rumors regarding enterprises or commercial services, risk of information leak through a company’s own employees, and other risks related to false advertising, pharmaceutical law and truth-in-advertising laws.
- In cases where urgent information is detected, we provide emergency notification and consulting on how to respond.