February 16, 2018

Automatic detection of sensitive content

Gang Wang

Yian Gao

Recommended Citation

Wang, Gang and Gao, Yian, "Automatic detection of sensitive content", Technical Disclosure Commons, (February 16, 2018)
http://www.tdcommons.org/dpubs_series/1055

Automatic detection of sensitive content

ABSTRACT

This disclosure describes techniques to automatically identify text content is inappropriate or likely in violation of organizational policy. A user interface is provided that enables users to make suitable corrections to written material. The described techniques can be implemented in software (e.g., in document authoring or communications applications).

KEYWORDS

- Document authoring
- Communication policy
- Email
- Inappropriate content
- Sensitive content

BACKGROUND

It is important that content shared with other users in professional environments such as workplaces adhere to organizational policies regarding content. For example, organizational communications policies often specify words, phrases, sentences, and other content that are impermissible in documents on organization computers, networks, and other systems. However, authors may inadvertently include inappropriate or sensitive words, phrases, or sentences in content and subsequently shared within or outside the organization.

Organizations and individuals can face difficulties and severe consequences if inappropriate words, phrases, or sentences are used in official documents or communication. For example, violation of some organizational policies can lead to disciplinary action against the respective authors. If such documents are shared outside the organization, or made public, the organization can also face public relations issues and/or legal challenges. It is thus important for organizations that documents, communications, and other content adhere to organizational policies.

Organizations rely on employee training to reduce the likelihood that inappropriate words/phrases are not used in written documents. However, employees may not always remember the training or to adhere to guidelines. It is burdensome and costly for organizations to devote resources for manual review of documents to ensure that organizational policies are not violated. Manual review can also be error-prone.

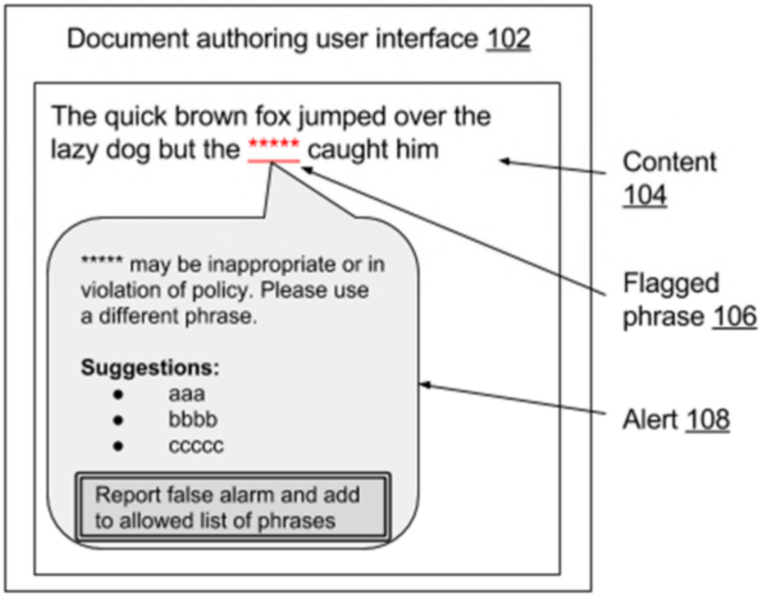
DESCRIPTION

This disclosure describes techniques that can be implemented in document and email authoring software applications to automatically inspect documents and flag inappropriate or sensitive words, phrases, and sentences. The techniques are implemented with user permission and express consent to perform automatic analysis of text entered by users. Content authors are alerted by flagging the identified text and suitable alternatives are recommended that are acceptable and appropriate.

To predict inappropriate or sensitive content, a built-in repository of such words, phrases, and sentences is included. The repository includes inappropriate or sensitive content in the language that is generally applicable to the country and/or culture of the author and organization. Further, a configuration model is incorporated to enable customization of the repository based on factors such as organizational policies, laws and regulations, and cultural and/or social customs.

Customization can include adding sensitive or inappropriate words, phrases or sentences to the built-in repository or re-designating existing content in the repository as not inappropriate and sensitive. The repository can also include information that provides reasons for the designation of particular words or phrases as inappropriate or sensitive. These details can also be customized. The repository can also include suggested alternatives for content that is flagged, and the suggested alternatives can be customized.

A user interface is provided to automatically highlight through visual cues inappropriate or sensitive words, phrases and sentences in document or communication content. For example, objectionable text may be underlined or highlighted using a specific format/color. The user interface enables authors to report that the flagged language is appropriate in the given context, and to indicate that the highlighted phrase be re-designated as appropriate.



Document authoring user interface 102
Content 104
Flagged phrase 106
Alert 108
Report false alarm and add to allowed list of phrases

The quick brown fox jumped over the lazy dog but the ***** caught him

***** may be inappropriate or in violation of policy. Please use a different phrase.

Suggestions:

- aaa
- bbbb
- ccccc

Fig. 1: Detection of inappropriate and sensitive content

Fig. 1 illustrates a document authoring user interface (102) with content (104) that includes objectionable text "*****" (106) shown in red font and underlined in red color. Interaction with, e.g., right-clicking or selecting the highlighted text, displays an alert (108) that includes a reason to flag the text as sensitive or inappropriate content and suggests acceptable and suggested alternatives. The user interface enables users to report false alarms when the flagged language is appropriate in the given context.

The techniques can optionally use machine learning to predict whether a given word, phrase or sentence and similar as well as related words, phrases and sentences are appropriate and acceptable within a specific country, culture, corporation and/or organization. When users provide consent, machine learning models can learn suitability of particular words or phrases based on configuration of the techniques in particular user contexts, and user reports obtained from the user interface.

Different organizational contexts may include different interpretations of appropriateness of particular words or phrases. For example, medical information of individuals may be configured as inappropriate in most organizational contexts. However, documents with medical information may be deemed appropriate when shared within a group of authorized medical professionals in a healthcare organization. In another example, discussions regarding undergarments are typically disallowed within most organizations. However, if the organization is a manufacturer of undergarments, words related to undergarments, e.g., color, texture, or material of undergarments, are not flagged in this context. In this example, the machine learning model learns that “garment color,” “garment texture,” “garment material,” and “garment style” concepts are appropriate for the organization in the context of discussions related to the garment and not a person that wears the garment. Based on this knowledge, “undergarment texture,” “undergarment material,” “undergarment style,” and “undergarment color” are predicted to be acceptable in the context of an organization that is a manufacturer of undergarments.

Optionally, confirmation is obtained from an administrator or other authority in the organization that terms such as “undergarment texture,” “undergarment material,” “undergarment style” and/or “undergarment color” are acceptable within the context of documents and communications related to undergarment products. The described techniques can thus be augmented or enhanced with machine learning to predict inappropriate and insensitive words/phrases/sentences and to reduce instances of false alarms.

The present techniques reduce the likelihood that inappropriate and sensitive words, phrases or sentences are inadvertently included in organizational communications, and can help reduce legal and/or public relations risks for organizations. Document preparation and communications (e.g., email, messaging/chat, etc.) and content authoring software applications can incorporate the described techniques.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques to automatically identify text content is inappropriate or likely in violation of organizational policy. A user interface is provided that enables users to make suitable corrections to written material. The described techniques can be implemented in software (e.g., in document authoring or communications applications).