Technical Disclosure Commons

Defensive Publications Series

November 2019

Automatic content filtering in virtual assistants for kids

Yuzhao Ni

Automatic content filtering in virtual assistants for kids

ABSTRACT

Virtual assistant responses need to be both useful and safe for kids and families. However, this is currently not always the case. For example, virtual assistant responses can sometimes unexpectedly include explicit answers or answers that are otherwise unsuitable for kids. However, restricting searches can prevent the virtual assistant from surfacing useful, family-safe responses. There is no systematic way to filter non-textual media content, e.g., music, video, etc. Per the techniques of this disclosure, a library of content classifiers is provided that filters out various categories of content inappropriate for children, e.g., explicit content, violent content, etc. A query to a virtual assistant and the responses to the query are filtered by the classifiers. Depending on the context, e.g., current audience, the virtual assistant surfaces filtered responses to queries.

KEYWORDS

- Content filtering
- Content classification
- Virtual assistant
- Adult content
- Family context
- Kids mode
- Family mode
- Content rating

BACKGROUND

Virtual assistant responses need to be both useful and safe for kids and families. However, this is currently not always the case. For example, virtual assistant responses can sometimes unexpectedly include explicit answers or answers that are otherwise unsuitable for kids. However, restricting searches can prevent the virtual assistant from surfacing useful, family-safe responses. There is no systematic way to filter non-textual media content, e.g., music, video, etc.

DESCRIPTION

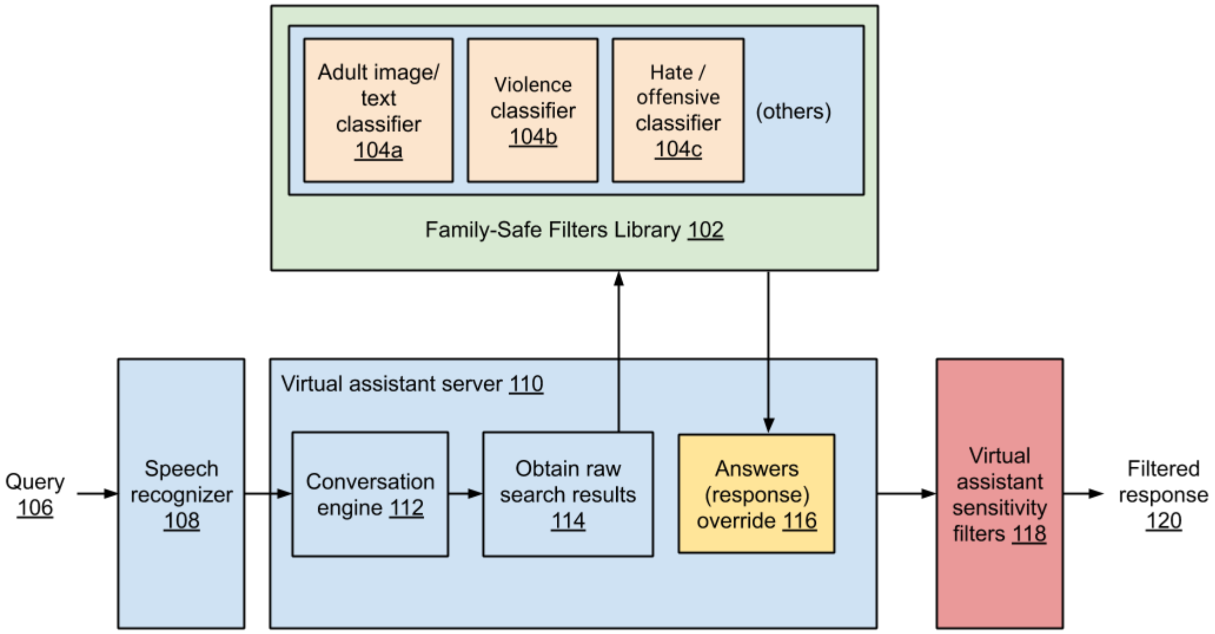
Adult image/ text classifier 104a
Violence classifier 104b
Hate / offensive classifier 104c
(others)
Family-Safe Filters Library 102
Query 106
Speech recognizer 108
Virtual assistant server 110
Conversation engine 112
Obtain raw search results 114
Answers (response) override 116
Virtual assistant sensitivity filters 118
Filtered response 120

Fig. 1: Content filtering in virtual assistants for kids

This disclosure describes techniques to filter virtual assistant queries and responses so as to make them child-safe. The techniques are illustrated in Fig. 1. A library (102) of classifiers or filters is maintained that can detect various categories of content deemed inappropriate to children, e.g., content that includes adult images or text (104a), portrays violence (104b), is hateful or offensive (104c), is dangerous or illegal, etc. A query (106) is transcribed by a speech recognizer (108) and sent to a virtual assistant server (110). The actions of the virtual assistant server depend on the context (determined with user permission), which includes information about the current audience of the virtual assistant, whether the virtual assistant is in kids mode, etc.

The virtual assistant server includes a conversation engine (112) that organizes the query-response stream. The conversation engine passes the transcribed query to a unit (114) that obtains raw search results, e.g., by invoking a search engine. Both the query and the search results are sent to the classifier library. The filters of the classifier library filter out inappropriate content, e.g., content that is determined to have an inappropriateness/sensitivity score that is above certain thresholds. The filters return search responses that have inappropriateness score below threshold.

Under certain conditions, e.g., if no filtered response is sufficiently relevant, or if the query partially or fully matches certain prohibited phrases, then an answer override unit (116) overrides the filtered responses with standard messages. The filtered search results are further sent to native sensitivity filters (118) of the virtual assistant and after applying the native sensitivity filters, final results are surfaced to the user (120).

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

Per the techniques of this disclosure, a library of content classifiers is provided that filters out various categories of content inappropriate for children, e.g., explicit content, violent content, etc. A query to a virtual assistant and the responses to the query are filtered by the classifiers. Depending on the context, e.g., current audience, the virtual assistant surfaces filtered responses to queries.

REFERENCES

[1] Dinosearch, "Kiddle - visual search engine for kids," https://www.kiddle.co/ accessed on Oct. 22, 2019.