July 2023

Determining User Journey Risk Trajectories in Information Seeking Sessions

Abhishek Roy

Ellie Jin

Rebecca Umbach

Determining User Journey Risk Trajectories in Information Seeking Sessions

ABSTRACT

This disclosure describes techniques to measure risk trajectories in user journeys involving online search tasks performed by a user using a search engine, chatbot, or another query answering engine. Search tasks include the user interacting with (e.g., selecting and viewing) search results, chatbot generated answers, and web pages linked to those search results. Based on metadata about user-submitted queries, the user search session is divided into user visit segments that include sensitive queries by the user relating to seeking assistance ("help seeking") or seeking potentially detrimental content ("harm seeking"). Determination of risk categories for sensitive queries are made (e.g., by a human evaluator and/or automated system) and a risk trajectory for the user is determined over a user session based on determined risk valuations. The user session is categorized based on risk trajectory to determine potential of risk for harm seeking by the user. Described techniques can measure risk trajectories that include multiple interactions of a user journey and enable improvement in providing assistance to help-seeking and harm-seeking users. The discussion in this paper is the result of exploratory studies conducted to assess risks associated with user journeys - using mental health as a particular example.

KEYWORDS

- Search engine
- Chatbot
- Query answering
- Query intent
- Sensitive query
- Self-harm
- Information seeking session
- User journey
- Risk trajectory
- Suicide

BACKGROUND

Efforts to ensure satisfying and safe user experiences in search engines, chatbots, or other applications can include gating (or restricting) access to harmful content for users and providing helpful interventions that aid in well-being for users who are looking for help online. User queries in searches that are related to harm-seeking or help-seeking can be considered sensitive search queries since they may be indicative of a user's mental health or well-being. In some examples, a user who inputs a query related to self-harm may be struggling with mental health issues related to that type of act; a user who inputs a query of "where to buy drugs" may be struggling with addiction; or a user who searches for "how to get help with depression" may be struggling with depression or a related condition.

For physical or offline actions, practitioners such as criminologists, psychologists, and psychiatrists have diagnostic tools to ascertain the risks of a person being a danger to themselves or to society. However, there are few or no tools that allow automated determination or estimation of risks in users' online behavior, e.g., when performing searches using a search engine or seeking answers from a chatbot. Also, previous ways of estimating user experiences in such applications do not take into account the actions a user may take in sequence during a session or across multiple sessions, which can lend added context and affect the evaluation of risk of the user's behavior being missed.

DESCRIPTION

This disclosure describes techniques that enable risks of self-harm as evidenced by users' online behavior to be ascertained by adapting established diagnostic tools from the offline world to online user journeys. The techniques enable evaluation of online behavior in interaction with search engines, chatbots, etc. to obtain a deeper understanding of user experience, assess impact

of interventions on harm-seeking user journeys, and monitor user experience for sensitive query spaces. The techniques can be deployed in contexts where user surveys or experiments are not feasible. The discussion in this paper is the result of exploratory studies conducted to assess risks associated with user journeys - using mental health as a particular example.

The described techniques measure severity of risk in user behavior by evaluating user queries, applying ratings to the queries, and determining a trajectory of user ratings over a user session. The techniques, implemented with user permission, can help determine whether users move from help seeking to harm seeking and the factors that may trigger or influence that change. The determined risk trajectories can assist in understanding a user journey by evaluating queries over time.

A search engine as referred to herein may be a general purpose internet search engine, a special-purpose or domain specific search engine, a search interface within an application, a chatbot, or any other modality through which a user can perform a search or seek answers to queries. The described techniques can be implemented on any suitable device or system, e.g., desktop or laptop computer, portable user device (e.g., a smartphone), server device(s), etc.

User risk trajectories in search tasks

The described techniques include automated evaluation of sensitive queries input by a user to determine risk trajectories for user journeys. Sensitive queries include help-seeking queries from users (e.g., hotline/counseling seeking, symptom management, self-diagnosis, substance use disorder treatment, etc.) and harm-seeking queries from users (e.g., methods of self-harm, criminal behavior, substance abuse, etc.).

A risk rating scale can be used to assign different values to different categories of harm-seeking, e.g., indicating a level of severity, or a probability or potentiality that users will perform

a self-harming act. For example, a rating scale similar to the Columbia-Suicide Severity Rating Scale can be used to assign severity values to various self-harm triggering queries. In one example, a risk rating scale for potential self-harm in persons can include hopelessness, with a rating of 1 (e.g., “no one cares I am having a panic attack”); a wish to be harmed, with a rating of 2; vague thoughts of doing self-harm without mention of specific methods, with a rating of 3; seeking a method of self-harm, with a rating of 4, and self-harm being imminent, with a rating of 5.

Help-seeking queries that the user performs are also assigned values for different levels of severity of help-seeking in a help rating scale. Help-seeking queries can be assigned negative values that are the opposite of positive harm-seeking values (or vice-versa). For example, neutral queries can be assigned 0, which are queries that are non-harm-seeking, non-help-seeking, and informational only (e.g., “flight to city x”, searches for friends and family). Queries related to assessment of the user’s symptoms can be assigned a value of -1. Queries related to self-help reduction (e.g., “how to stop having a panic attack”) can also be assigned a value of -1. Queries that are requests to connect with external help (e.g., a help hotline or counseling) can be assigned a value of -2. Other values can be used for these or other types of queries, and/or for other types of risks of the user (e.g., addiction and substance abuse, performing violent acts, etc.).

Severity values can be associated with individual queries of multiple successive queries input by the user to form a risk trajectory for the user. An example of a risk trajectory of a user based on use of a search engine for a search task is shown in Fig. 1.

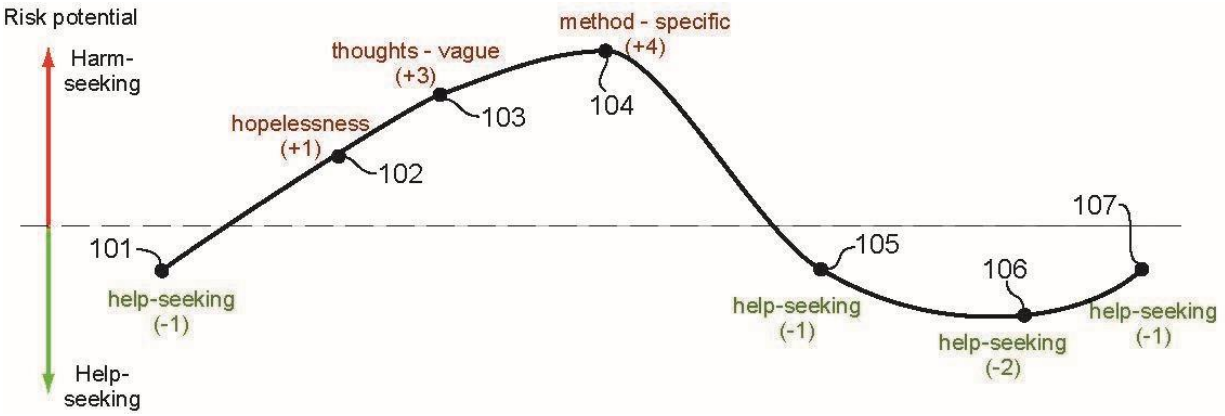

Fig. 1: Example risk trajectory

As shown in Fig. 1, the user journey can include successive queries input to a search engine over a user visit segment, e.g., from left to right in Fig. 1. The queries can be evaluated, categorized, and assigned severity values, and are ordered based on the times they were submitted. In this example, the user queries start with a help-seeking query (101), and this is followed by several harm-seeking queries of successively greater severity values (102, 103, 104) as categorized using the rating scale described above. The user then returns to submitting help-seeking queries (105, 106, 107). For example, help-seeking query 106 can be a search to connect with external help (e.g., using a hotline link) as described above. The help-seeking query 107 is the last query of this example visit segment as determined based on, for example, the user stopping to submit queries, viewing a web page for longer than a threshold period of time, or submitting further search queries that are not sensitive queries.

Determining risk trajectories and risk categories for users

An example method to determine user risk trajectories and risk categories from use of a search engine is shown in Fig. 2.

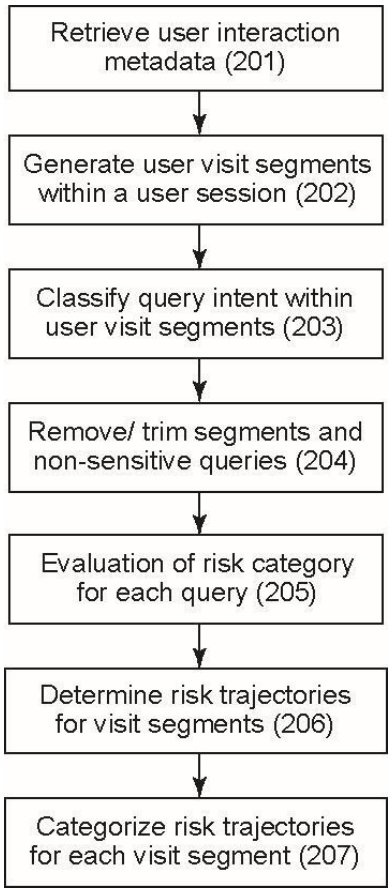

Fig. 2: Determining user risk trajectories and risk categories from search tasks

As shown in Fig. 2, user interaction metadata is retrieved (201). The metadata can be retrieved from logs and can include queries that were input by a user, as well as other metadata that describe events, actions that were performed by the user during use of the search engine or for pages selected via search results (e.g., clicks, selections, etc.) or chatbot answers, and timestamps indicative of when the actions were performed. The metadata can include labels assigned to queries by classifiers that can categorize the subjects or topics of queries based on, for example, keywords in the queries. Queries can be labeled with multiple categories or classifications of different specificity or breadth by a classifier, e.g., a machine learning model trained for query classification.

The metadata that is retrieved includes at least one query that is a sensitive query. Sensitive queries indicate a particular query intent of the user that is related to a sensitive subject, including harm-seeking queries and help-seeking queries. For example, a sensitive query intent may be indicated by labels assigned to the queries.

The retrieved metadata can be limited to a particular user session, e.g., a single login session of the user, or a set of query-related actions that occur within a certain time period (e.g., three hours). For example, the retrieved metadata can be associated with a sequence of all user events that occurred in a random sample of daily sessions, with at least one sensitive query included in each session.

User visit segments within the user session are generated based on the retrieved metadata (202). A user visit segment can be an uninterrupted series of user actions which include a search task involving one or more search queries. Timestamps associated with the user actions of the user session are examined to determine when interruptions occur in the user session. Interruptions are indicated by a threshold time period in which no user actions occurred, and user visit segments are defined between such interruptions. In some cases, user actions within the visit segments can be sequenced according to their timestamps, if the user actions are not already provided sequentially in the metadata.

Query intents are classified within the generated user visit segments (203). The classifications of the queries may be included in the metadata as described above or can be determined by a classifier. Sensitive queries and non-sensitive queries are determined as classifications of query intent.

User visit segments that do not have at least one sensitive query are removed, and user visit segments are trimmed of non-sensitive queries (204). In some examples, the first sensitive

query and the last sensitive query of a visit segment are detected, and the queries and user actions occurring before and after the first and last sensitive queries are trimmed. Intermediate queries between the first and last sensitive queries can be retained. In some other cases, intermediate non-sensitive queries can be trimmed.

In other cases, all or some non-sensitive queries may be retained in the user visit segments to be evaluated and/or to deduce patterns of user behavior. For example, intermediate non-sensitive queries can be used in an evaluation of whether these queries distracted the user or altered a risk trajectory of the user. Non-sensitive queries that precede or succeed sensitive queries, and the topics of such queries, can also be retained for the evaluation. In some cases, if the labels assigned to queries do not have a minimum confidence of accuracy, all queries and actions can be retained since there may be less certainty regarding which queries are sensitive queries.

Next, evaluation of risk categories of the queries in the user visit segments (205) is performed. For example, such evaluation can be performed by human(s) that have experience in evaluating and detecting particular types of harm-seeking and help-seeking in users based on submitted text such as queries. The evaluator can assign a severity value (rating) to each query, e.g., similar to the examples described above, based on the category determined for the query. In some cases, some types of sensitive queries can be determined and rated by a system, e.g., based on detection of particular keywords in queries such as “hotline,” “helpline,” or “prevention.”

A risk trajectory is determined for each visit segment (206). The risk trajectory provides a sequence of risk severity ratings over time based on the timestamps of the queries, as in the example of Fig. 1 described above. The risk trajectories are categorized for each visit segment (207) to provide risk categories. A number of risk categories can be defined for different types of

risk trajectories, e.g., based on how the severity values of a trajectory change or fluctuate. In some examples, risk trajectories can be categorized in one of four categories: flat, backsliding, escalating, and de-escalating, which are based on the directions or trends of changes of severity values within the risk trajectory.

In a flat trajectory, the user-issued queries are at the same level of severity or have severity values that do not change significantly (e.g., more than 1 level) throughout the visit segment. In a backsliding trajectory, one of two patterns are evident: the user issued one or more significant help-seeking queries such as requests to connect with external help (e.g., a help hotline or counseling), and then issued harm-seeking queries; or the user issued one or more severe harm-seeking queries, then one or more harm-seeking queries and/or help-seeking queries that were less severe, and then returned to issuing severe harm-seeking queries.

In an escalating trajectory, the user issued help-seeking or less-severe harm-seeking queries and then issued only severe harm-seeking queries. In a de-escalating trajectory, the user issued at least one severe harm-seeking query and then issued less-severe and/or help seeking queries.

In some cases, an overall risk trajectory can be determined and categorized similarly as described above that describes multiple evaluated user visit segments of a user session, e.g., if these visit segments occurred successively and close in time to each other. A risk category determined for a risk trajectory can be provided to human evaluators and/or to automated systems that can provide functions or assistance to the user, e.g., in the case of trajectories indicative of greater risk of harm to the user.

Determined trajectories allow better detection and evaluation of risks in user behavior. For example, the tracking of queries in a risk trajectory provides improved understanding of

queries which may appear to be non-sensitive, or at least less severe when evaluated individually but may actually indicate significantly harmful potential when evaluated holistically, e.g., in a sequence or otherwise as part of a risk trajectory (e.g., a user issuing a query asking for "tallest waterfalls in area A" after a query of "is it painful to drown").

The determined user trajectories can be used as a benchmarking tool for assessing impact of intervention on sensitive user journeys. This can be useful to understand the efficacy of a new intervention or other help function provided by the system that provides search or query-answering functions, such as a coping module, or for ranking changes as related to specific results. Determining a percentage of users who are escalating vs. de-escalating in their journeys, or backsliding, can allow correlation of risk trajectory changes to product changes. A search engine or other query-answering engine can make use of described features to monitor product and feature impact on users and enable improvement of user interface experiences for users on sensitive search journeys.

In some cases, user friction can also be measured in the user visit segments described above. User friction is the incremental effort of users to find satisfying search results using the search engine. For example, user friction can be estimated by counting friction actions by the user within the segment, such as returning to a search results page after selecting a particular result and viewing a landing page, refining the search by modifying the query, scrolling the search results page or a landing page, selecting additional links within the landing page and later accessed pages, opening a tab in a browser or other application for a new search, enabling a particular search mode (for text, images, etc.), etc. A high friction score can be used in evaluation of a risk trajectory, e.g., determining whether measured amounts of friction have an effect on or correlation with particular risk trajectories. User friction measured for different risk

trajectories can be compared and evaluated to determine effects of friction on the risk trajectories.

In addition, search tasks can be classified as having a completion status of "completed" or "abandoned," where a completed search task ends with the user selecting and viewing help information, website, or other information from the results for more than a threshold amount of time, and an abandoned search task is otherwise indicated. Completion statuses can be correlated with user journey friction to determine correlations and effects on user trajectories.

The various features of the system are implemented only with user permission to access user information that serves as input to the system (e.g., user queries, user-provided images or other content items, user context information, camera input, user's preferences, etc.). Users 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, and if the user is sent content or communications from a server. 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. 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 measure risk trajectories in user journeys involving online search tasks performed by a user using a search engine, chatbot, or another query answering engine. Search tasks include the user interacting with (e.g., selecting and viewing) search results, chatbot generated answers, and web pages linked to those search results. Based on metadata about user-submitted queries, the user search session is divided into user visit

segments that include sensitive queries by the user relating to seeking assistance (“help seeking”) or seeking potentially detrimental content (“harm seeking”). Determination of risk categories for sensitive queries are made (e.g., by a human evaluator and/or automated system) and a risk trajectory for the user is determined over a user session based on determined risk valuations. The user session is categorized based on risk trajectory to determine potential of risk for harm seeking by the user. Described techniques can measure risk trajectories that include multiple interactions of a user journey and enable improvement in providing assistance to help-seeking and harm-seeking users. The discussion in this paper is the result of exploratory studies conducted to assess risks associated with user journeys - using mental health as a particular example.