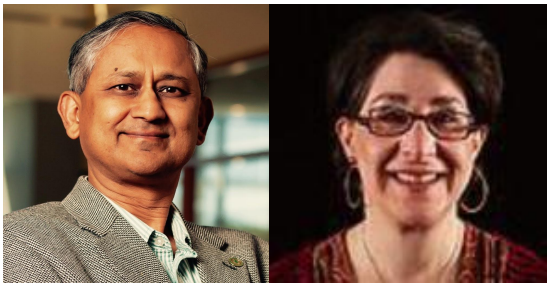
COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Amit Sheth (University of South Carolina) & Valerie Shalin (Wright State University), June 2022



Title: *Semantic analysis of social media and new Big Data to understanding COVID-19's impact on mental health, addiction, and gender-based violence*

YouTube Recording with Slides

June 2022 CIC Webinar Information

Transcript Editor: Julie Meunier

Transcript

Amit Sheth:
Slide 1
D'accord, très bien. Nous allons discuter de notre analyse des problèmes de santé publique. Lorsque nous analysons une très grande quantité de contenu sur les médias sociaux, l'accent est mis sur l'analyse de la santé mentale, des problèmes liés à la dépendance et ce que nous appelons l'indice de qualité sociale. L'Institut d'IA de l'Université de Caroline du Sud, l'Université d'État de Géorgie et l'Université Wright State sont des participants ou des membres de l'équipe appartenant à ces institutions.

Slide 2
Nous connaissons tous l'impact massif de cette pandémie.

Slide 3
Et cela a également eu un impact significatif sur la dépression, l'anxiété, d'autres problèmes de santé mentale, les addictions et l'abus de substances.

Slide 4
Nous nous sommes demandés si nous pouvions comprendre comment diverses décisions politiques, politiques sociales, économiques et de santé publique, ainsi que les choix faits par le gouvernement et les décideurs, affectent le bien-être de la société - les membres de la société. Notre boîte à outils est l'analyse des médias sociaux améliorée par la connaissance.

Slide 5
Ainsi, nous la présentons comme suit : nous avons collecté des données entre mars 2020 et fin juillet à janvier 2021. Cela couvrait deux grandes vagues de COVID-19. Nous avions 25 termes de recherche. Nous avons collecté 12 milliards de tweets. En plus de ces mégadonnées, nous disposions de nombreuses autres données à utiliser. Cela fait partie essentielle de l'étude car ce n'est pas seulement l'analyse des médias sociaux. Les données ne suffisent pas, il faut beaucoup d'autres informations pertinentes pour analyser les données. Nous disposions d'informations liées à la localisation, nous avons examiné des sous-forums spécifiques pour former nos modèles linguistiques, et nous avions des bases

de connaissances liées à la santé mentale, à l'ontologie de l'abus de drogue du DSM-V, DAO, DBpedia, Wikidata, UMLS. De plus, nous avons dû examiner les événements sur le terrain. Par exemple, les aides aux prêts, les aides au revenu des ménages, les politiques de dépistage, toutes ces décisions prises par les dirigeants des États ou les décideurs fédéraux. Nous devons également comprendre ce qui s'est passé quand et où. Et il y a d'autres événements que nous devons aussi comprendre.

Slide 6
Notre objectif était de comprendre le contenu sur les médias sociaux lié à la santé mentale. Vous pouvez voir quelques exemples ici.

Slide 7
Et nous avons développé une mesure empirique que nous appelons l'indice de qualité sociale. Il agrège les composantes de la santé mentale et les composantes de la dépendance et de l'abus de substances. Nous analysons les données des médias sociaux pour ces composantes afin de créer un indice de qualité sociale.

Slide 8
L'infrastructure pour une analyse de ce type est assez complète. Vous avez les données Twitter, mais vous avez aussi des articles de presse. Je mentionnerai rapidement pourquoi les articles de presse sont importants - diverses bases de connaissances. Ensuite, vous devez effectuer toute une série d'analyses, vous pouvez le voir dans le deuxième encadré vertical, puis cela passe aux modèles linguistiques, aux modèles de sujets, aux cartographies sémantiques, pour une compréhension plus approfondie. Et notre équipe travaille sur l'apprentissage enrichi par la connaissance, nous avons donc des méthodes d'apprentissage profond améliorées qui utilisent la connaissance du domaine pour mieux comprendre le langage. À partir de là, pour comprendre le contenu lié à la dépression, à l'anxiété, à la dépendance, à l'abus de substances, tout cela dans le contexte de la COVID, ce qui conduit à des calculs d'indice de qualité sociale, il y avait des composantes de formation de modèle linguistique, et tout cela n'est pas montré ici dans cette image.

Slide 9
Alors, quelles sont les innovations ? Nous avons utilisé les actualités pour identifier en continu de nouvelles entités liées à la COVID. Et dans ce contexte, nous utilisons l'extraction de la localisation en utilisant l'ontologie Geonames, l'API Open Street Map et d'autres éléments. Nous avons utilisé de multiples vocabulaires et graphes de connaissances pour l'extraction sémantique, nous avons utilisé un corpus de sous-forums spécialisés pour la formation des modèles linguistiques et des modèles de sujets. Et nous avons formé des classificateurs pour mettre à l'échelle l'analyse des mégadonnées. Nous ne pouvions pas faire l'analyse - pour analyser de telles quantités de données, nous devions créer ces classificateurs et ensuite comprendre la dépression, la dépendance, l'anxiété et d'autres problèmes. Avec cette base, je vais laisser la parole à Valerie pour expliquer ce que nous avons découvert.

Valerie Shalin:
Slide 10
Merci, Amit. Permettez-moi de vous donner une idée des types d'analyses rendus possibles grâce aux capacités décrites par Amit. Voici donc quelques exemples de graphiques d'état montrant les taux de COVID et l'indice de qualité sociale au fil du temps. Comme Brandon [Johnson], nous pensons que le temps est absolument crucial ici. Les taux de COVID par habitant sont sur la gauche, l'indice de qualité sociale est sur l'axe de droite. Le haut est mauvais, le bas est bon. Le temps est sur l'axe des x et il y a une discontinuité marquée par les croisillons. Je ne m'attends pas à ce que vous lisiez attentivement ces

graphiques, sauf que vous remarquerez certainement qu'il y a très peu de corrélation entre l'indice de qualité sociale que nous avons identifié et la prévalence de la COVID. Il y a un endroit - oups, revenons une diapositive en arrière, juste pour une seconde - où vous remarquerez une corrélation. Et c'est vers la fin, et nous pensons que cela reflète probablement une variable latente. Il s'agit de la saison des fêtes d'hiver, et probablement ce qui se passe, c'est que les fêtes d'hiver augmentent à la fois les taux de COVID, et nous savons déjà que les fêtes d'hiver sont très éprouvantes pour la santé mentale. Cela a donc beaucoup de sens et cela sert de validation informelle de la métrique de l'indice de qualité sociale. Diapositive suivante.

Amit:
Je ne peux pas avancer, je ne sais pas ce que - Lauren, je ne peux pas...

Lauren Close:
Voyons, je pourrais peut-être partager pour vous.

Florence Hudson:
Avez-vous essayé de cliquer simplement sur la diapositive ou d'utiliser les flèches vers le haut ou vers le bas ? Avez-vous essayé plusieurs options différentes ? Ok, ça y est.

Valerie:
Slide 11
D'accord, l'un des points que nous voulons souligner, c'est que les vacances, bien sûr, ne sont pas le seul événement que nous pouvons maintenant examiner. Et voici un graphique pour l'État de Washington. Ce que nous vous montrons, ce sont les changements de politique que nous avons identifiés dans les actualités, alignés sur les différentes périodes que nous avons marquées avec les lignes verticales. Notre point, c'est qu'il est désormais possible de rechercher des modèles liant cet indice de qualité sociale à la mise en œuvre de la politique par l'État. Diapositive suivante, Amit.

Slide 12
Donc l'indice de qualité sociale est une mesure standardisée. Il est indépendant de la population de l'État ou de la quantité d'activité sur les médias sociaux, et parce qu'il est standardisé de cette manière, nous pouvons comparer les États. Et nous l'avons fait - diapositive suivante.

Slide 13
Certaines analyses de regroupement par État pour voir si nous pouvons discerner des modèles. Et vous pouvez voir que le Connecticut, la Louisiane, le New Jersey, le Nevada, etc. semblent tous suivre le même genre de modèle. L'indice de qualité sociale est mauvais au début, s'améliore, puis se détériore. Cela concerne une période de temps à haute résolution tronquée. Et un groupe différent : la Floride, la Géorgie, le Michigan, etc. ont un modèle SQI pire - SQI pire. Donc nous pouvons dévoiler certains de ces modèles au fil du temps. Bien sûr, nous devons faire davantage d'analyses longitudinales et de séries temporelles pour dégager les causes potentielles, mais il semble que ce sont les problèmes économiques qui seront le facteur principal ici. Diapositive suivante.

Slide 14
Alors, qu'est-ce qui est maintenant possible avec les outils développés par l'équipe de Caroline du Sud ? Nous avons une analyse pilotée par la connaissance des médias sociaux à un niveau abstrait et cela fournit un ensemble de données à très haute résolution, tant en termes de temps que d'espace. C'est rapide et cela permet d'obtenir des mesures séparables de la santé mentale qui ne sont pas réalisables

dans les enquêtes classiques. Nous pouvons obtenir beaucoup de données réparties sur de grandes parties de l'espace et du temps, et c'est le type de données dont vous avez besoin pour éliminer les confusions et commencer à évaluer si vos décisions politiques ont un impact sur votre population. Il s'agit donc d'un outil de recherche très important qui renforce nos capacités en science, mais plus largement, il permet une surveillance en temps réel et une préparation à l'atténuation. Et bien sûr, il soutient les décideurs en général. Cela conclut notre présentation.