Kap Code

CONCEPTION D'UN ALGORITHME PERMETTANT DE DÉTECTER L'HÉSITATION VACCINALE ANTI-PAPILLOMAVIRUS HUMAIN AU SEIN DE MESSAGES ISSUS DES RÉSEAUX SOCIAUX

Pierre Foulquié[1], Anaïs Gedik[1], Simon Renner[1], Paméla Voillot[1], Adel Mebarki[1] et Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France

Detec't

INTRODUCTION

La France est un des pays où l'hésitation vaccinale anti-papillomavirus humain (HPV) est la plus forte au monde. Cette hésitation s'observe particulièrement sur les réseaux sociaux, interface ou les internautes peuvent s'exprimer librement sur leur santé. L'amélioration de l'acceptabilité vaccinale HPV passe par la compréhension des déterminants de l'hésitation. **Un algorithme d'analyse sémantique capable d'identifier les messages exprimés sur les réseaux sociaux contenant une hésitation vaccinale anti-HPV permettrait d'analyser et de comprendre ce phénomène.**

MATÉRIEL ET MÉTHODES

Un corpus de messages associés à la vaccination anti-HPV, postés entre 2006 et 2019, a été extrait dans le cadre du projet Detec't [1] à partir de **17 sources francophones**. Les 23 mots-clés d'extraction évoquaient plusieurs sujets associés à la sphère du papillomavirus : **la vaccination anti-HPV, la sexualité et l'anatomie**. Une annotation d'un échantillon du corpus a été effectuée par 3 annotateurs qui disposaient d'une charte d'annotation basée sur la définition de l'hésitation vaccinale de l'OMS [2]. Elle a permis de classer les messages comme exprimant de l'hésitation ou non et d'extraire des expressions des différentes perceptions vaccinales (anti-vaccin, pro-vaccin). Le gold standard (GS) ainsi créé a été réparti en **2 jeux de données**. Un premier jeu, dit « d'entraînement » et contenant 85% des données, a été utilisé pour entraîner le modèle. Le deuxième jeu, désigné comme jeu « de validation » et constitué des 15% restants, a servi à la validation de la méthode.

À partir du jeu d'entrainement, plusieurs variables ont été déterminées à l'aide des formes syntaxiques des messages (N-grams), de la présence des mots de champs lexicaux spécifiques (anti-vaccin, pro-vaccin, etc.) et du *word embedding* (représentation contextuelle des mots via un modèle Glove [3]).

Par la suite et afin de mettre en place un modèle performant en termes de précision de détection d'hésitation vaccinale, une recherche de la meilleure combinaison entre différents classifieurs (*support vector classification*, *logistic regression*, *random forest*, et *extreme gradient boosting*) et les différentes variables identifiées précédemment a été effectuée.

RÉSULTATS

1 370 messages contenant une mention de vaccination anti-HPV ont été extraits pour annotation. Les sources les plus présentes dans l'échantillon étaient les forums Doctissimo et Santé Médecine, avec respectivement 615 et 193 messages, suivis de Twitter (n= 395). Le terme « Gardasil » est le mot d'extraction ayant permis de recueillir le plus de messages dans cet échantillon (n= 967).

L'annotation a permis d'identifier 497 messages (36%) représentant une hésitation vaccinale et 891 (64%) une perception positive, négative ou neutre de la vaccination. Les jeux de données d'entrainement et de validation étaient composés respectivement de 1 164 et 206 messages. La meilleure combinaison de variables identifiée est constituée des 300 premiers N-grams en terme de pouvoir prédictif de l'hésitation vaccinale. Ces formes syntaxiques sont composées d'un mot unique ou d'associations de deux ou trois mots telles que « savoir » ; « mère veut »; « rapports sexuels ».

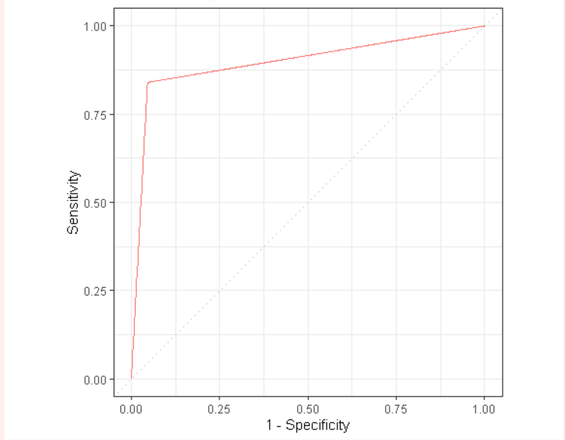
Suite aux tests, le meilleur modèle choisi est un classifieur binaire *Random Forest*, discriminant correctement l'hésitation des autres perceptions dans 91% des cas. Les performances du modèle, calculées à partir de la matrice de confusion (Tableau 1), sont regroupées dans le Tableau 2. Les messages classés comme évocateur d'hésitation vaccinale exprimaient une réelle hésitation dans 91 % des cas (valeur prédictive positive). Parmi tous les messages annotés comme exprimant une hésitation vaccinale, 84 % des messages ont été identifiés par notre modèle (sensibilité). Le Graphique 1 présente la courbe ROC du modèle.

		Référence	
		oui	non
Prédiction	oui	63	6
	non	12	125

Tableau 1 - Matrice de confusion du modèle

Précision	Sensibilité	Spécificité	Valeur prédite positive	Valeur prédite négative
91,26%	84,00%	95,42%	91,30%	91,24%

Tableau 2 – Performances du modèle



Graphique 1 – Courbe ROC du modèle

CONCLUSION

Développer un algorithme d'analyse sémantique capable d'identifier une hésitation vaccinale anti-HPV au sein de messages issus des réseaux sociaux pourrait se révéler être un **nouvel outil d'aide à l'identification des déterminants de l'hésitation dans le cadre d'une couverture insuffisante.**

Les performances de l'algorithme sur des données n'ayant pas servi au développement de ce dernier démontre que ce type d'outil est efficace pour identifier puis analyser les messages d'internautes exprimant une hésitation vaccinale anti-HPV.

Ce travail ouvre de nombreuses possibilités de travaux futurs. Des méthodes complémentaires, permettant par exemple d'identifier des causalités, pourraient permettre d'identifier les facteurs de cette hésitation. De plus, l'étude des utilisateurs exprimant une hésitation vaccinale permettrait d'établir des profils types et d'étudier l'évolution temporelle de cette hésitation. Ceci pourrait ouvrir la voie à l'instauration d'outils de monitoring de la vaccination sur les réseaux sociaux dans un objectif de santé publique.

RÉFÉRENCES

[1] Abdellaoui, R., Schück, S., Texier, N., & Burgun, A. (2017). Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help?. JMIR public health and surveillance, 3(2), e36.

[2] Ten threats to global health in 2019, https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

[3] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).