Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210308

Fuzzy Matching for Symptom Detection in Tweets: Application to Covid-19 During the First Wave of the Pandemic in France

Carole FAVIEZ[a,1] Pierre FOULQUIÉ[b], Xiaoyi CHEN[a], Adel MEBARKI[b], Sophie QUENNELLE[a,c], Nathalie TEXIER[b], Sandrine KATSAHIAN[a,d], Stéphane SCHUCK[b], and Anita BURGUN[a,e,f]

[a] *Centre de Recherche des Cordeliers, Sorbonne Université, INSERM, Université de Paris, F-75006, Paris, France*
[b] *Kap Code, Paris, France*
[c] *M3C-Necker, Hôpital Necker-Enfants Malades, AP-HP, F-75015, Paris, France*
[d] *Hôpital européen Georges Pompidou, Unité d'épidémiologie et de recherche clinique, AP-HP, F-75015, Paris, France*
[e] *Hôpital Necker-Enfants Malades, Département d'informatique médicale, AP-HP, F-75015, Paris, France*
[f] *PaRis Artificial Intelligence Research InstitutE (PRAIRIE), France*

Abstract. The exhaustive automatic detection of symptoms in social media posts is made difficult by the presence of colloquial expressions, misspellings and inflected forms of words. The detection of self-reported symptoms is of major importance for emergent diseases like the Covid-19. In this study, we aimed to (1) develop an algorithm based on fuzzy matching to detect symptoms in tweets, (2) establish a comprehensive list of Covid-19-related symptoms and (3) evaluate the fuzzy matching for Covid-19-related symptom detection in French tweets. The Covid-19-related symptom list was built based on the aggregation of different data sources. French Covid-19-related tweets were automatically extracted using a dedicated data broker during the first wave of the pandemic in France. The fuzzy matching parameters were finetuned using all symptoms from MedDRA and then evaluated on a subset of 5000 Covid-19-related tweets in French for the detection of symptoms from our Covid-19-related list. The fuzzy matching improved the detection by the addition of 42% more correct matches with an 81% precision.

Keywords. Content analysis, social media, fuzzy matching, symptoms, Covid-19

1. Introduction

Information retrieved from social media can provide insights for various health-related studies [1], including symptom detection [2]. However, the comprehensive automatic information extraction from social media such as Twitter is challenging due to the presence of colloquial expressions, inflected forms of words and misspellings.

[1] Corresponding Author, Carole Faviez, Centre de Recherche des Cordeliers, INSERM, 15 rue de l'école de médecine, F-75006, Paris, France; E-mail: carole.faviez@inserm.fr.

Algorithms such as fuzzy matching which address this issue have to be carefully parametrized to restrict the generation of extra noise.

During the unprecedented Covid-19 pandemic, due to the absence of prior knowledge, patient experiences have been intensively discussed on social media. At the same time, a list of symptoms was progressively established: firstly respiratory and general disorders (e.g., cough, fever), lately some digestive, neurological, mental health symptoms (e.g., diarrhea, anosmia, insomnia), and even “Kawasaki-like” symptoms among children. It is thus of great interest to analyze self-reported Covid-19-related symptoms and explore the emergence of symptoms in social media.

Our objectives are to (1) develop a fuzzy matching algorithm to detect symptoms in tweets, (2) establish a comprehensive list of Covid-19-related symptoms and (3) evaluate the fuzzy matching for Covid-19-related symptom detection in French tweets.

2. Methods

2.1. Fuzzy matching

Spelling mistakes are frequent in Social media. Moreover, some languages, such as French, have many inflected forms (e.g., “paralysie” (paralysis), “paralyser” (to paralyze), “paralysé” (paralyzed)). To extend the detection of symptoms in French tweets by taking into account misspelling and inflected forms, we developed a fuzzy matching algorithm, which allows the mapping between a term present in a thesaurus (a *thesaurus term*) and a word found in a tweet that conveys a similar meaning even if the two entities are not exactly the same. The Levenshtein distance [3] was computed between *thesaurus terms* and variants from the tweets. Two types of parameters were finetuned:

- The proportion of accepted edits between the *thesaurus term* and the compared segment from the tweets using the Levenshtein distance
- The length of the prefix, namely, the number of characters at the beginning of a term that have to remain identical

The fuzzy matching settings were adapted for French symptom recognition considering all symptoms from the Medical Dictionary for Regulatory Activities (MedDRA) in order to maximize the number of detectable terms and the proportion of correct matches (precision). The fuzzy matching performances were evaluated in the context of Covid-19-related symptoms in French tweets.

2.2. Covid-19-related message extraction

The analysis period was chosen as from January 1, 2020 to May 11, 2020 (the first lockdown lift) to cover the first wave of the pandemic in France. French Covid-19-related tweets were targeted based on a list of keywords from three categories: Covid-19 (e.g., covid, coronavirus), transmission control measures (e.g., “confinement” (lockdown), “restezchezvous” (stayathome)) and Covid symptoms (e.g., covid+, “apresj20” (afterd20)), and obtained through a data broker as part of Kap Code Detec’t solution [4]. This extraction pipeline is limited in volume to 20% of the set of targeted posts per day. The extracted tweets went through several preprocessing steps (conversion to lowercase and removal of URLs, accents, diacritics, punctuation and abusive spaces between words).

2.3. Covid-19 related symptoms – constitution of the list and detection in tweets

An initial list of specific Covid-19-related symptoms was manually established, based on four sources: (1) lists published by state or global health organizations such as the WHO[2] or the NIH[3], (2) symptoms published in early literature (from January 2020 to July 2020), (3) symptoms described in specialized media reports, such as specific dermatologic symptoms, (4) lists established by clinicians, especially by a pediatric cardiologist (SQ) for a complete description of "Kawasaki-like" symptoms in children.

This list was reviewed by public health experts (AB, SK). MedDRA version 23.0 was used as core terminology for symptom definition and identification. Consequently, symptoms from our initial list were manually mapped to at least one Preferred Term (PT) from MedDRA. This list was then completed by the addition of all symptoms from the Covid-19-related lexicon created by Sarker et al. [5]. This lexicon was made of symptoms extracted from Covid-19-related tweets in English that were mapped to UMLS IDs. When possible, we automatically converted UMLS IDs to MedDRA's PTs. Otherwise, we manually searched for a corresponding PT in MedDRA. The resulting list of PT labels was considered as our list of Covid-19-related symptoms.

In order to minimize the chances of missing symptoms expressed in messages, we enriched our set of PT terms by synonymous terms and colloquial expressions. This enrichment was performed following a three-step process. (1) We added from the MedDRA hierarchy all Low Level Terms (LLT) related to the selected PTs to our list as synonyms. (2) We translated all terms and expressions identified by Sarker et al. to French and added them as synonyms of the related PTs. (3) We manually added common expressions and colloquial language as synonyms based on our experience of medical concept enrichment for pharmacovigilance monitoring [6,7]. A final check of all added synonyms was performed: some synonyms were moved from one PT to another, and ambiguous synonyms that were expected to be noisy were deleted.

All Covid-19-related symptoms (PTs and their synonyms) were applied the same preprocessing as for tweets (conversion to lowercase, removal of accents, etc.). Results were aggregated at the PT level.

3. Results

3.1. Finetuning of the fuzzy matching algorithm

The fuzzy matching algorithm was tuned with different settings for all MedDRA symptoms detection in a set of 5000 tweets. For each test, we compared the performance of fuzzy matching to exact matching. A variant matched by the fuzzy matching was considered as a correct match if it consisted in a correct variation of the initial term and was not prone to cause additional noise. When the number of variants identified by the fuzzy matching was beyond 120, a random sample of 120 variants was reviewed for precision assessment. The results are summarized in table 1.

[2] https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19

[3] https://covid19.nih.gov

Table 1. Parametrization of the fuzzy matching (FM) algorithm on a first set of 5000 tweets

Iteration	Prefix length	Maximal distance (%)	Number of exact matches	Additional FM terms	FM precision
1	3	50%	267	30006	2%
2	3	33%	267	1185	21%
3	3	25%	267	446	34%
4	3	10%	267	13	**100%**
5	5	25%	267	177	87%
6	5	10%	267	13	**100%**

To balance between coverage and precision, prefix length = 5 characters and maximal distance = 25% were used for the following analyses. Consequently, the fuzzy matching was not applied on words with 5 characters or less.

3.2. Covid-19 related symptoms detection

Our final list of Covid-19-related symptoms contained 223 different PTs associated with 3113 synonyms (88% of LLTs in MedDRA and 12% of manual enrichments). The list is available upon request from the corresponding author (CF).

The Covid-19-related corpus was constituted of 792,193 French tweets extracted between January 1, 2020 and May 11, 2020 and containing at least one of the predefined keywords. 48,382 tweets (6%) contained at least one symptom (PTs and synonyms) from our list of Covid-19-related symptoms. 57,406 occurrences of symptoms were detected among these tweets, including 17,348 additional terms identified thanks to the fuzzy matching. Most frequent identified symptoms corresponded mostly to frequent Covid-19-related symptoms (e.g., Fatigue, severe acute respiratory syndrome, cough, pyrexia) as well as anxiety-related symptoms (e.g., fear, panic attack, depression, anxiety).

3.3. Evaluation of the fuzzy matching algorithm in the context of Covid-19

The fuzzy matching finetuned for all MedDRA symptoms was evaluated for Covid-19-related symptoms. Performances were assessed on a random subset of the Covid-19-related corpus of 5000 tweets. In this evaluation subset, 313 Covid-19-related symptoms were identified in 261 messages when considering the exact match, while using fuzzy matching, 474 Covid-related symptoms were identified (+51%) in 362 messages. The 161 supplementary variants identified through fuzzy matching were reviewed. Among the 161 generated forms, 131 were correct (precision of 81%), which means that the fuzzy matching enabled the detection of 42% supplementary correct terms. The 30 erroneous variants corresponded to 7 distinct PTs. Among them, variants for the three PTs *constipation* (e.g., “constitution” instead of “constipation”), *muscle contracture* (e.g., “contracte” instead of “contracture”) and *tenderness* (“sensibilise” instead of “sensibilité”) represented 87% (26/30) of errors.

4. Discussion

In this article, we developed a fuzzy matching algorithm for symptom detection in social media and evaluated it for Covid-19 in French tweets. The fuzzy matching improved substantially the performance of the Covid-19-related symptom detection with reasonable amount of noise. While most of the work using fuzzy matching do not report on explicit evidence regarding the choice of the parameters in comparison with exact matching, we assessed the impact of different values on precision. We obtained similar precision as the data-centric system developed by Sarker et al. to detect drug names in tweets in English (precision between 79 - 84%) [8]. Compared to data driven methods, our approach requires more manual steps: post-validation is recommended to check main errors and further optimize the precision. However, compared to data driven models, e.g., [9,10], lexicon-based methods take advantage of multilingual reference terminology thus can be easily converted to other languages. It would be interesting to develop a hybrid approach combining these two ideas to improve the performance for unknown symptoms. In terms of results, our list of 223 different symptoms validated by experts is to our knowledge the most comprehensive, covering most of the symptoms identified from English tweets in previous studies (36 symptoms [9] and 51 symptoms [11]). The uncovered ones were rare and often less specific to Covid-19 (e.g., hair loss, weight gain). As perspective, we intend to take advantage of this work to evaluate the capacity of early detection of Covid-19-related symptoms on social media in French and English.

References

[1] Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a Tool for Health Research: A Systematic Review. Am J Public Health. janv 2017;107(1):e1-8.

[2] Yazdavar AH, Al-Olimat HS, Ebrahimi M, Bajaj G, Banerjee T, Thirunarayan K, et al. Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. Proc IEEE ACM Int Conf Adv Soc Netw Anal Min. août 2017;2017:1191-8.

[3] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady. 1 févr 1966;10:707.

[4] Abdellaoui R, Schück S, Texier N, Burgun A. 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 Surveill. 22 juin 2017;3(2):e36.

[5] Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang Y-C. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J Am Med Inform Assoc. 1 août 2020;27(8):1310-5.

[6] Chen X, Deldossi M, Aboukhamis R, Faviez C, Dahamna B, Karapetiantz P, et al. Mining Adverse Drug Reactions in Social Media with Named Entity Recognition and Semantic Methods. Stud Health Technol Inform. 2017;245:322-6.

[7] Kürzinger M-L, Schück S, Texier N, Abdellaoui R, Faviez C, Pouget J, et al. Web-Based Signal Detection Using Medical Forums Data in France: Comparative Analysis. J Med Internet Res. 20 nov 2018;20(11):e10466.

[8] Sarker A, Gonzalez-Hernandez G. An unsupervised and customizable misspelling generator for mining noisy health-related text sources. J Biomed Inform. déc 2018;88:98-107.

[9] Guo J-W, Radloff CL, Wawrzynski SE, Cloyes KG. Mining twitter to explore the emergence of COVID-19 symptoms. Public Health Nurs. nov 2020;37(6):934-40.

[10] Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, et al. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. JMIR Public Health Surveill. 8 juin 2020;6(2):e19509.

[11] Jeon J, Baruah G, Sarabadani S, Palanica A. Identification of Risk Factors and Symptoms of COVID-19: Analysis of Biomedical Literature and Social Media Data. J Med Internet Res. 2 oct 2020;22(10):e20509.