

تحليل العوامل النفسية والاجتماعية في مجتمع Suomi24 للسرطان عبر الإنترنت.

Psychosocial Factor Analysis In Suomi24 Online Cancer Community

مراد أو صالح¹، مؤمن إبراهيم²¹University of Oulu, Faculty of Information Technology and Electrical Engineering

PO Box 4500, 90014 Oulu, (FINLAND)

Mourad.Oussalah@oulu.fi,²University of Oulu, Faculty of Information Technology and Electrical Engineering ,

PO Box 4500, 90014 Oulu, (FINLAND)

oussalahmourad42@gmail.com

تاريخ النشر: جوان/2025

تاريخ القبول: 2025/05/29

تاريخ الإرسال: 2025/05/25

ملخص

يعد Suomi24 أكبر وسائل التواصل الاجتماعي الفنلندية التي تركز على المواضيع، ويوفر فرصاً ممتازة لدراسة تفاعلات المستخدمين مع مختلف مواضيع النقاش المعاصرة. تركز هذه الورقة البحثية على مجتمع السرطان في Suomi24. باستخدام أدوات معالجة اللغة الطبيعية المبتكرة التي تستفيد من أنطولوجيا جديدة لمطابقة موضوعات النقاش الفنلندية حول السرطان مع المعلومات السياقية المرتبطة بها وتحليل التكرار المشترك القائم على المسافة متعددة المستويات، والذي يأخذ في الاعتبار العلاقات النحوية، يتم طرح تصنيف جديد للعوامل النفسية. يبدأ هذا التصنيف من خلال صياغة مجموعة أولية من الرموز يدوياً لكل فئة، ثم يتم توسيعها من خلال استكشاف العلاقة الدلالية والمعلومات السياقية.

وقد كشف التحليل التصنيفي عن خمسة عوامل نفسية اجتماعية رئيسية، تتكون حسب ترتيب أهميتها من الجوانب الاجتماعية والمالية والعلاجية والمرضية، والجوانب المتعلقة بالوفاة. وهذا يعزز بعض الدراسات السابقة في الدراسات السريرية التي تؤكد على هيمنة العوامل الاجتماعية. من ناحية أخرى، تم طرح منصة مفتوحة وإتاحتها لمساعدة كل من الباحثين والأطباء السريريين على استكشاف وتصور النتائج من مجموعة بيانات Suomi24. يمهد هذا العمل الطريق لمزيد من توصيف المجتمعات الصحية على الإنترنت على نطاق واسع.

الكلمات المفتاحية: التقييم في النصوص، علم الأنطولوجيا، الصحة، العوامل النفسية والاجتماعية

Abstract

Suomi24 is Finland's largest topic-centric social media, providing excellent opportunities to study users' interactions with various contemporary discussion topics. This paper focuses on cancer community in Suomi24. Using innovative natural language processing tools that makes use of a new ontology for matching Finnish cancer discussion topics with the associated contextual information and a multi-level distance-based co-occurrence analysis, which takes into account syntactic relations, a new categorization of psychological factors is put forward. This categorization starts by manually crafting an initial set of tokens for each category, which are then expanded by exploring the semantic relationship and contextual information.

The categorization analysis revealed five main psychosocial factors, which consist, in the order of their importance, of Social, Finance, Treatment, Illness and Death related aspects. This reinforces some previous studies in clinical studies, which stress on the dominance of the social factors. On the other hand, an open platform has been put forward and made available to help both researchers and clinicians to explore and visualize the findings from the Suomi24 dataset. This work paves the way for further characterization of online health communities at wide.

Keywords: Text mining, ontology, health, psychosocial factors.

1. مدخل:

إن الزيادة الهائلة في وسائل التواصل الاجتماعي على كل مستويات الفئات السكانية والتخصصات تمهد الطريق لاتجاهات بحثية جديدة في العلوم السلوكية وتكنولوجيا المعلومات والعلوم الصحية. في هذا الأخير، تم الاعتراف على نطاق واسع بوسائل التواصل الاجتماعي من خلال قدرتها على نقل الدعم الاجتماعي للمرضى وأقاربهم/أصدقائهم. على سبيل المثال، استُخدمت وسائل التواصل الاجتماعي لتحليل المشاعر [26]، وفهم تصور المرضى لموضوع معين من قبل المرضى [38]، واستخراج المشاعر الاجتماعية من المحتوى النصي [4]، ورسم خرائط الأدلة المتعلقة بالانتحار [25] والاكتئاب [3، 23، 28]، والمشاعر [34]، وغيرها. في هذا الصدد، يتجلى الدعم الاجتماعي في المقام الأول من خلال أربع فئات متميزة [31-32]: (1) الدعم العاطفي، الذي يساعد المرضى على تلبية احتياجاتهم العاطفية أو العاطفية؛ (2) الدعم التقديري الذي يعزز ثقة المريض في التغلب على التحديات الكامنة؛ الدعم المعلوماتي، الذي يزود المريض بحيل غير متوقعة أو مفيدة من تجارب الآخرين والمجتمع؛ الدعم الشبكي الذي يذكر المريض بالمساعدات المحتملة التي يمكن أن يحصل عليها من المجتمع الذي ينتمي إليه. وقد تم تحديد هذه النتائج من خلال مسح محتوى منصات التواصل الاجتماعي الرئيسية؛ وهي فيسبوك وإنستغرام وتويتر وسناب شات.

(1) موقع Suomi24^[33] هو أكبر موقع تواصل اجتماعي في فنلندا و يركز على المواضيع، كما أنه أحد أكبر منتديات النقاش على الإنترنت في العالم يستعمل لغة غير الإنجليزية^[12]. ووفقاً لموقع TNS metrix (2013)¹، يزور 86% من الفنلنديين الذين يستخدمون الإنترنت الموقع كل شهر - على الأقل لمعرفة إجابة عن سؤال بحثوا عنه في جوجل. تنشأ شعبية المنتدى من عدم الكشف عن هوية المحادثات، مما يحفز المستخدمين على الانخراط بحرية في مواضيع النقاش المختلفة التي تشمل الاتجاهات السياسية والسياسات والقضايا الصحية والأخبار والهوايات. يفتح هذا الأمر فرصاً للتطوير المنهجي وتحليل البيانات وأدوات التصور من أجل توصيف المنشورات بالإضافة إلى ديناميكياتها السياقية والاجتماعية والزمنية. وبشكل أكثر تحديداً، تغطي مجموعة بيانات Suomi24 عقدين من تاريخ المناقشات على الإنترنت وتحتوي على أكثر من ملياري رمز من المناقشات الفنلندية مع ما يقرب من 832.000 مستخدم فريد أسبوعياً يقضي كل منهم في المتوسط 7 دقائق على المنتدى في كل مرة. وهذا يجعل من مجموعة بيانات Suomi24 مختبراً ممتازاً لاستكشاف الجمع بين الخصائص اللغوية للنص والسمات الإحصائية لديناميكيات المناقشة مع أسئلة أبحاث العلوم الاجتماعية، مما يفتح آفاقاً فريدة للجمع بين الاتجاهات النوعية والديناميكيات الكمية للتفاعل وعواطف المستخدمين ومشاعرهم، أو استكشاف جوانب مختلفة من الأدلة المتضاربة. ومن الناحية الأكاديمية، يوفر هذا الأمر ميزة واضحة مقارنةً بوسائل التواصل الاجتماعي الأخرى، مثل فيسبوك وتويتر، التي لا يمكن عادةً الحصول على مجموعة كاملة من البيانات لمثل هذا الفحص النقدي. وهذا ما يحفز الاستخدام الحالي لمجموعة البيانات للتحليل الشبيه بالصحة لاستكشاف العوامل النفسية والاجتماعية لمرضى السرطان.

من المعروف أن المناقشات المجهولة على الإنترنت توفر ساحة مهمة لمعالجة المشاكل المتعلقة بالصحة ومشاركتها^[10، 31، 32] حيث تحظى كفاءة العلاج والآثار الجانبية والتجارب الشخصية بشعبية كبيرة. وبالفعل، تتم الإشارة إلى الأدوية وجرعاتها وآثارها الجانبية وتوافرها، على سبيل المثال، عند مناقشة تشخيص الأمراض وتطورها. إلى جانب ذلك، غالباً ما يرتبط انتقاد ممارسات الرعاية الصحية بمحدودية أو عدم كفاية الأدوية أو الإفراط في تناولها وإضفاء الطابع الطبي العلني على حياة الناس. وبشكل عام، تفتح بيانات وسائل التواصل الاجتماعي المجهولة الهوية منظوراً للحديث بين الأقران حول ما قد يُعتقد أنه من المحرمات في القضايا الصحية، متجاوزةً الإطار المهني المعتاد للعلاقة بين الطبيب والمريض. من هذا المنظور، يمكن النظر إلى الحديث عن الأدوية ونمط الحياة المرتبط بها كجزء من الساحة التي يتصور فيها الناس مخاوفهم الصحية ويتفاوضون بشأنها مع مدى ملاءمة الممارسات والقيود الحالية. وغالباً ما يتم تشخيص مرضى السرطان المهددة حياتهم بالقلق والاكتئاب، الأمر الذي يتطلب اهتماماً خاصاً من المجتمع ككل^[18]. غالباً ما يكون المنتدى مكاناً ممتازاً لنقل هذا الدعم وتبسيط الضوء عليه

حيث يحاول المرضى أيضاً توسيع نطاق معرفتهم بالسرطان والتحقق من صحة/دحض المعلومات التي يتلقونها من أخصائي الرعاية الصحية. وقد حدث هذا على نطاق واسع في تهديدات المناقشة في سومي24 وفقاً لتحقيقات سابقة^[13]. ويشمل ذلك الدعم العاطفي من خلال التعبيرات المختلفة عن الرعاية والاهتمامات التي تساعد في نهاية المطاف على تحسين مزاج المريض، ومشاركة الخبرات والطرق الجديدة التي قد تزيد من وعي المريض وتقديره وشبكة علاقاته الاجتماعية بطريقة تعزز الثقة مع تعزيز العلاقة (الجديدة) القائمة على السمات المشتركة. وبعبارة فضفاضة، يساهم هذا أيضاً، إلى حد ما، في تمكين مجتمع مرضى السرطان عبر الإنترنت.

وتتراوح الأبحاث حول المدونات بين تحديد المدونات^[29]، وتحديد مصدر إعادة نشر المدونات^[2]، وتحليل الروابط^[35]، وتحليل الطنين^[37]، وتحليل الشبكات الاجتماعية^[1]. كما تم تطوير عدد من الأدوات للمساعدة في تصور جميع أنواع بيانات المدونات، مع التركيز بشكل أساسي على تصور الروابط بين المدونات والمدونين الأفراد، أو ديناميكية مصطلحات الطنين، أو الموضوعات أو السمات المختارة بمرور الوقت، ومحتوى السببية للمدونات، من بين أمور أخرى، انظر، على سبيل المثال،^[14].

على الرغم من وجود تاريخ طويل في تحليل وتصور واستكشاف المناقشات الإنجليزية عبر الإنترنت باستخدام أساليب التنقيب عن النصوص والتعلم الآلي^[9، 11-12]، إلا أن التعامل الفعال مع مجموعة بيانات Suomi24 بعيد كل البعد عن أن يكون مهمة سهلة وتم تحديد العديد من التحديات. أولاً، على الرغم من أن المنتدى يشجع المناقشة القائمة على المواضيع، إلا أنه غالباً ما تحدث محادثة متعلقة بالصحة أو السرطان في مواضيع غير متعلقة بالصحة. ثانياً، عدم الكشف عن الهوية وغياب مفهوم معرف المستخدم يجعل أي تحليل شبيه بالشبكة صعباً ومحدوداً للغاية. ثالثاً، استخدام أدوات معالجة اللغة الطبيعية القياسية محدود للغاية بالنسبة للغة الفنلندية. في الواقع، في حين أن حجم المفردات في اللغة الإنجليزية يميل نحو منحني لوغاريتمي مع نمو حجم المادة، فمن المعتاد في اللغة الفنلندية أنه مع نمو حجم مجموعة البيانات، يستمر عدد أشكال الكلمات في النمو خطياً. يؤدي هذا إلى ندرة البيانات، وهي مشكلة بالنسبة للعديد من طرق التعلم الآلي التي تعتمد على سبيل المثال على تمثيلات تشبه حقيقية الكلمات للبيانات. علاوة على ذلك، في النصوص العامية، التي تعج بالأخطاء الإملائية والمختصرات والكلمات المخترعة، تكون فائدة التحليل اللغوي محدودة. رابعاً، غالباً ما يسفر استخدام واجهة برمجة تطبيقات الترجمة مفتوحة المصدر عن استجابات غير مرضية بسبب محدوديتها المتأصلة في كل من النطاق والتراكيب. خامساً، يعد السرطان من أكثر القضايا الحساسة التي لا يشعر المرضى عموماً بالراحة في الكتابة عنها بالتفصيل، مما يضيف صعوبة إضافية للباحثين في اكتشاف القرائن المفيدة، وربما تحديد المقاييس المناسبة ذات الصلة بالأطباء السريريين.

تقترح الدراسة الحالية أداة عملية يمكن استخدامها لاستكشاف مجموعة بيانات Suomi24 لتحديد الأنماط المحتملة المرتبطة بالاستفسارات الصحية. يتم التحقيق بشكل خاص في العوامل النفسية والاجتماعية لمرضى السرطان. ولذلك، فإن الأسئلة البحثية الرئيسية التي تم تسليط الضوء عليها في هذه المقالة هي :

(1) كيف يمكن تحديد العوامل النفسية والاجتماعية من منتديات المناقشة عبر الإنترنت؟ .

(2) كيفية تصميم نهج آلي يحدد هذه العوامل في Suomi24؟.

يعالج الحل المقترح في هذه الورقة البحثية الأسئلة البحثية المذكورة أعلاه التي تقتصر على خطاب مرضى السرطان باستخدام مزيج من تقنيات معالجة اللغة الطبيعية والتحليل الإحصائي والملاحظات الإحصائية السليمة. على وجه الخصوص، تم ابتكار منهجيات أصلية وتنفيذها من أجل (1) جمع مجموعة البيانات ذات الصلة من مجموعة بيانات Suomi24 الشاملة، (2) تحديد فئات العوامل النفسية والاجتماعية الرئيسية، (3) توصيف كل فئة من خلال أنطولوجيا محددة، (4) تقييم مدى كل عامل نفسي اجتماعي في المجموعة، (5) التفاعل مع مجتمع أوسع من خلال بوابة مخصصة لذلك. ومع ذلك، نعتقد أن مثل هذا النهج المطور يمكن استخدامه في نهاية المطاف لاستكشاف أنماط اجتماعية لغوية أخرى من مجموعة بيانات Suomi24. تم تنظيم بقية المقال على النحو التالي. يفصل القسم 2 من هذه الورقة البحثية إجراءات جمع البيانات، والسمات الرئيسية لمجموعة البيانات ومنهجية المعالجة المسبقة. يتناول القسم 3 تحليل البيانات، والذي يتضمن تحليل مجموعة البيانات وتحديد العوامل النفسية والاجتماعية. يسلط القسم 4 الضوء على تنفيذ المنصة. وأخيراً، نختتم المقال ونسلط الضوء على العمل المستقبلي.

2. الأساليب:

1.2 سومي 24 والرعاية الصحية:

Suomi24 هو منتدى نقاش يتكون من مناقشات متاحة للجمهور من إنشاء المستخدمين يتم تجميعها حسب المحتوى، مثل الترفيه والهوايات والسفر والصحة. وكونه مجهول الهوية، يمكن للمستخدمين بدء مناقشاتهم أو المساهمة في المناقشات القائمة. في هذه الدراسة، تم استرجاع بيانات المنتدى من قاعدة بيانات منظمة، تم استيعابها من قبل مزود الخدمة. ويمنح ترخيص قاعدة البيانات، وفقاً لاتفاقيات حقوق الطبع والنشر من قبل المنظمة العالمية للملكية الفكرية، الحق في استخدام ونسخ مجموعة البيانات لأغراض تعليمية وتدرسية وبحثية [33]. تحتوي مجموعة البيانات على 352,725 منشوراً في فئة الصحة مقسمة إلى 16 فئة فرعية، وهي "أسأل أسئلتك الصحية"، "تحديد النسل"، "تحديد النسل"، "الوفاة والحداد"، "الأمراض"، "المخدرات والإدمان"، "الصحة العامة"، "الرعاية الصحية"، "خدمات الرعاية الصحية"، "الأدوية"، "صحة الرجال"، "الصحة النفسية والعافية"، "صحة الفم"، "الجراحة التجميلية"،

"الحواس (الأعضاء الحسية)"، "التحكم في الوزن"، و"صحة المرأة". إن توزيع عدد الملاحظات بين الفئات الـ 16 ليس موحدًا، انظر الجدول 1. تحتوي العناوين والمشاركات على 2.6 و 75.9 كلمة في المتوسط، على التوالي. القيم الوسيطة لعدد الكلمات هي 2 للعناوين و 49 للمنشورات كما هو موضح في الجدول 1. من ناحية أخرى، حاولنا أيضًا تحديد مدى احتواء المنشورات على كلمات غير معروفة. يتم الحصول على هذا الأخير من خلال البحث عن وجود كلمة واحدة في القاموس الفنلندي أو إحدى قواعد بيانات الرسائل النصية القصيرة الفنلندية (على سبيل المثال، Slangi.net، Suomen murteiden sanakirja). ولذلك، إذا لم يتم العثور على كلمة ما في القاموس الفنلندي أو في قاعدتي بيانات الرسائل النصية القصيرة المذكورتين أعلاه، يتم تصنيفها على أنها غير معروفة في الجدول 1.

تكشف مراجعة سلاسل المناقشات المتعلقة بالصحة في Suomi24 الموضحة في الجدول 1 ما يلي:

- هناك طفرة في المناقشات المتعلقة بالأمراض والصحة النفسية/الصحة العقلية/الرفاهية كما يتضح من عدد كبير من المشاركات في هذه المواضيع.
- يستخدم المستخدمون الكتابة المطولة عندما يتعلق الأمر بقضايا الصحة النفسية، ربما لإظهار مشاعرهم وأحاسيسهم مع محاولة إيلاء اهتمام خاص بتجربة المريض ومناقشته. كما أنها تشير إلى الرغبة في مشاركة التجارب.
- وجد أن حدوث الصياغات المجهولة، على الرغم من ضآلتها، أكثر أهمية نسبيًا في موضوع الأمراض وموضوع الطب أكثر من غيرهما. ينشأ أحد تفسيرات هذه الملاحظة من المصطلحات التقنية المستخدمة للأمراض والطب، والتي أدت إلى أخطاء المستخدمين الإملائية الطوعية/غير الطوعية في الصياغة إلى وجود صياغات غير معروفة.
- على الرغم من أهميته، لم يتم ذكر "السرطان" كموضوع منفصل في منتدى المناقشة الصحية. وهذا ما يجعل البحث عن المناقشة المتعلقة بالسرطان موزعة على عدة مواضيع. على سبيل المثال، توجد الصياغات المتعلقة بالسرطان في مواضيع الأمراض والصحة العامة، والرعاية الصحية والصحة النفسية والرفاهية. هذا لا يعني بشكل بديهي أنه لا يمكن العثور على الصياغة المتعلقة بالسرطان في مكان آخر.

الجدول الأول. إحصائيات مجموعة بيانات SUOMI24 HEALTH

الفئة	#المشاركات	Mean	Median	متوسط %U words	إحصائيات الكلمات
اطرح أسئلتك الصحية	6,012	74.5	57	2.6	
تحديد النسل	13,213	58.6	47	2.4	
الوفاة والحداد	4,108	94.3	62	2.9	
الأمراض	86,035	76.4	52	4.8	
المخدرات ولادمانها	34,346	68.6	35	2.8	
الصحة العامة	16,415	69.2	49	2.1	
الرعاية الصحية	15,461	61.0	34	1.8	
خدمات الرعاية الصحية	236	61.5	35	1.7	
الأدوية	9,743	52.5	36	3.5	
صحة الرجل	4 4,067	55.8	36	2.3	
الصحة النفسية والرفاهية	70,017	103.0	66	1.8	
صحة الفم	10,959	59.3	43	1.9	
الجراحة التجميلية	6,123	55.6	39	3.7	
الأعضاء الحسية	8,164	64.5	46	2.1	
التحكم في الوزن	49,257	69.1	47	1.6	
صحة المرأة	18,569	69.5	55	1.7	

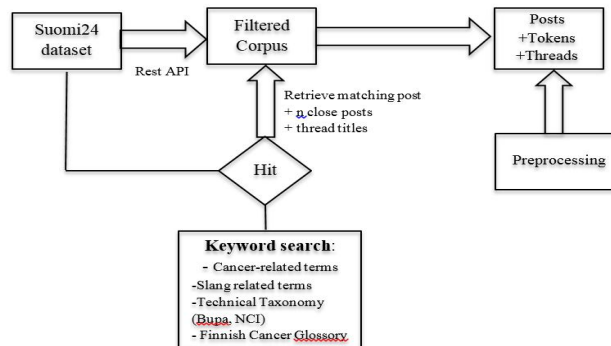
إن غياب المناقشة المتعلقة بالموضوع يسلط الضوء على أهمية البحث المتعلق بالكلمات الرئيسية عبر مجموعة بيانات Suomi24 الصحية بأكملها لتحديد الأنماط الرئيسية التي قد تكون مرتبطة بمرضى السرطان.

3. البيانات:

الخطوة الأولى في منهجيتنا هي جمع مجموعة البيانات المتعلقة بالسرطان من منتدى Suomi24. يوفر المنتدى واجهة برمجة تطبيقات للحصول على منشورات المنتدى من قاعدة البيانات. تم استخدام

طلب HTTP POST مع معلمة "كلمة رئيسية" لاسترداد المناقشات المقابلة. لهذا الغرض، اخترنا حقيقة من الكلمات الرئيسية التي تشمل:

1. السرطان ومرادفاته كما تم جمعها من القاموس الفنلندي على الإنترنت،
2. المصطلحات العامة المستخدمة للسرطان والأمراض المرتبطة بالسرطان التي تم جمعها من موقع slangui.net
3. قاموس طبي للأمراض المتعلقة بالسرطان مستخرج من موقع Bupa.co.uk (www.bupa.co.uk/health-information/cancer/cancer-terminology) مع ترجمته الفنلندية باستخدام واجهة برمجة تطبيقات الترجمة من Google
4. قاموس مصطلحات السرطان (8,334 مصطلحاً) من المعهد الوطني للسرطان التابع للمعاهد الوطنية للسرطان (NCI) والذي يغطي الأدوية وأنواع السرطان ومختلف أنواعه السريرية (مصطلحات عالمية، لذا لا حاجة للترجمة) ،
5. معجم مصطلحات السرطان الفنلندي الذي أنشأه البروفيسور تيبو ليلي، متاح في الرابط <https://www.kaikkisyovasta.fi/tietoa-syovasta/syopasanasto>
6. تنسيق مجموعة البيانات التي تم إرجاعها هو JSON، في حين تم استخدام مكتبة "ijson" Python لتكرار جميع البيانات المتاحة. لذلك، يكون إجراء اختيار مجموعة بيانات البحث على النحو التالي: أولاً، نأخذ في الاعتبار مجموعة البيانات المتعلقة بالصحة بأكملها المشار إليها في القسم 2.1 ونعتبرها مصدراً واحداً لمناقشة مرضى السرطان. ثانياً، نجري بحثاً نصياً عبر مجموعة بيانات Suomi24 بأكملها عن حقيقة الكلمات المفتاحية المذكورة أعلاه، وكلما تم العثور على تطابق لكلمة مفتاحية واحدة على الأقل، يتم استرجاع المنشور المرتبط واسم الموضوع وخمس مشاركات سابقة للموضوع المطابق وكذلك المشاركات الخمس التالية (شريطة ألا يكون المنشور في موضوع لاحق). إن الدافع وراء الرغبة في استرداد المشاركات السابقة واللاحقة للمشاركة المطابقة مع الموضوع الأساسي هو الحرص على النقاط سياق المناقشة. يتم عرض المخطط العام في الشكل 1.



الشكل 1. النهج العام لجمع البيانات

4. المعالجة المسبقة للبيانات:

تهدف مهمة المعالجة المسبقة إلى تحديد حدود المنشورات الفردية ومكوناتها المختلفة (الرموز) مع الأخذ في الاعتبار عددًا كبيرًا من المحتويات وأنماط الكتابة التي قد تكون مضللة والتي قد تتسبب في تجزئة غير مرضية. وهذا يشمل، على سبيل المثال، وجود روابط أو أحرف غير مألوفة أو مسافات في غير محلها أو أحرف ترقيم يمكن أن تؤثر على إجراء المزيد من التحليل. لهذا الغرض، تم تنفيذ المهام التالية لكل منشور في مجموعة بيانات السرطان.

1. يتم تحويل جميع النصوص إلى أحرف/رموز;
2. تتم إضافة مسافة بيضاء بعد كل نقطة إلا إذا لم تكن متوفرة;
3. يتم تحويل أي عدد من أحرف المسافات البيضاء المتتالية إلى حرف أبيض واحد;
4. تتم إزالة جميع عناوين URL مع أي حرف غير مألوف (مثل الأحرف الصينية واليونانية). كما تتم إزالة الأحرف غير المألوفة باستخدام وظائف regex;
5. تتم إزالة الكلمات الوقفية لتقليل قائمة الكلمات المفتاحية. تم استعارة هذه القائمة من UniNE من جامعة نوشاتيل^أ، بعد التدقيق اليدوي للإبقاء على جميع الصفات ومعلومات الوقت/التاريخ والمصطلحات المتعلقة بالنفي لأهميتها في الاستدلال اللاحق;
6. يتم تنفيذ الترميز لتحديد الرموز الفردية للمشاركات. لا يمكن لمحدد المسافة البيضاء الكلاسيكي أن يعمل في اللغة الفنلندية بسبب البنية المتأصلة في اللغة التي تجمع بين الكلمات الفردية لإنشاء كلمات أكثر طولاً. لذلك، فإن مساهمة محلل اللغة الفنلندية مهم في النهاية لتمييز الصياغة الفردية. لهذا الغرض، نستخدم المحلل اللغوي الفنلندي العميق^ب الذي تم تطويره في جامعة توركو;
7. تم اعتماد نهج الجذع البسيط. يهدف هذا إلى تحديد جذر/جذع الكلمة لتقليل عدد تصريفات الكلمات، والذي بدوره يقلل من العدد الإجمالي للكلمات/الرموز المستخدمة لتمثيل الجملة (ما بعد). لهذا الغرض، استخدمنا FinnPos^ج وهي مجموعة أدوات مفتوحة المصدر لوضع العلامات الصرفية والليمانية للغة الفنلندية.

5. لنتائج**1.5. التحليل الأولي:**

لفهم محتوى مجموعة البيانات التي تم جمعها، نقوم بالإبلاغ عن توزيع الكلمات الأكثر شيوعاً في المجموعة التي تمت تصفيتها وتكرار المصطلحات الأكثر حدوثاً في السرطان لتصور مدى التقارب/التكرار مع المصطلحات الأكثر حدوثاً. استُخدمت الخوارزمية 1 الموضحة في الشكل 2 لتحليل تواتر المصطلحات. الجدول 2 يظهر ملخصاً لسمات مجموعة البيانات.

الجدول الثاني. ملخص مجموعة البيانات

Size of dataset	400MB
Number of posts	55500
Number of threads	26052
Number of distinct words	41024
Time frame	2001-2015

في حين أن النتائج معروضة في الجدول 3.

الجدول الثالث. المصطلحات الأكثر شيوعاً والمصطلحات المتعلقة بالسرطان

Most repeated words		Most repeated cancer words	
Word	#number	Word	#number
saa	21862	määrä	3190
mun	7190	isä	1718
mies	7123	kuoli	1447
ihminen	6623	leikkaus	1324
apua	6505	raha	1197
euroa	6409	hoidon	1189
kokemuksia	5886	tauti	1039
tietää	5822	kasvain	345
päivä	5807	kaveri	248
oon	5793	tutkimus	179
nainen	5755	lääkitys	135
osaa	5311	hoito	107

Algorithm 1 repetition detection from Suomi24

Input finnishCancerKeywords, Suomi24 files, rightTopics=[Health], finnishStopwords, FreqDist from NLTK

Output graphs, mostRepeatedwords array, cancerRepeated array, outputFile

```

1: procedure REPETITION DETECTION
2:   for every file in Suomi24 files do
3:     items ← file/items
4:     for every item in items do
5:       topics ← item/topics
6:       for topic in topics do
7:         if topic in rightTopics then
8:           body ← item/body
9:           if body contains finnishCancerKey-
              words then
10:            Sentences ← Divide sentences in TranslatedTweet;
11:            Tokens ← Tokenize words in Sen-
              tences;
12:            totalDist ← FreqDist(words)
13:            end if
14:            end if
15:            end for
16:            end for
17:            end for
18:            for every item,value in totalDist do
19:              if item in finnishCancerKeywords then
20:                cancerRepeated ← [item,value]
21:              end if
22:            end for
23:            mostRepeatedwords ← totalDist[0:12]
24:            function VISUALIZE
25:              x-axis ← (mostRepeatedwords,cancerRepeated) items
26:              x-axis ← (mostRepeatedwords,cancerRepeated) values
27:              VISUALIZE
28:            end function
29: end procedure

```

الشكل 2. خوارزمية الكشف عن مصطلح السرطان المتكرر

وتعليقاً على هذه النتيجة، ينبغي للمرء أن يلاحظ ما يلي:

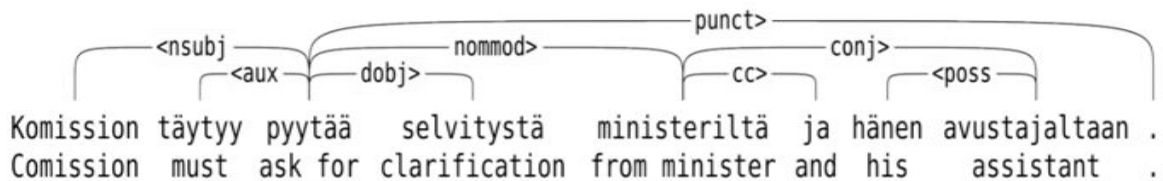
- تم الحصول على النتائج بعد إزالة كلمات التوقف، والتي من شأنها تجاهل المصطلحات الأقل فائدة في الدراسة.
- للتوضيح ، اقتصرنا على المصطلحات الاثني عشر الأكثر تكراراً في كل فئة (المصطلحات غير المتعلقة بالسرطان والمصطلحات المتعلقة بالسرطان). والدافع وراء ذلك أيضاً هو حقيقة أن المصطلحات المذكورة أعلاه تشكل نصف المصطلحات تقريباً التي لها ثلاثة تكرارات على الأقل في العينة المحددة.
- ترتبط غالبية المصطلحات غير المتعلقة بالسرطان بالمودة، وأنواع مختلفة من المساعدة (مشاركة الخبرة، والتمويل، والمشورة).
- يسلط الرسم البياني الضوء أيضاً على شبه هيمنة كلمة متعددة الأشكال ("saa")، والتي تترجم فعل "الحصول"، نظراً لاستخدامها الواسع في اللغة الفنلندية.
- وكما هو متوقع، فإن المصطلحات المتعلقة بالسرطان أقل تواتراً من المصطلحات غير المتعلقة بالسرطان بمعامل لا يقل عن اثنين. إلى جانب ذلك، فإن المصطلحات المتعلقة بالعلاج أقل تواتراً، وهو ما يحفز السياق المستخدم لتوليد المناقشات حول المصطلحات المتعلقة بالسرطان.
- يسلط خطاب السرطان الضوء على أهمية الجوانب المالية وآثار العلاج والوفاة والمرض والأدوية التي يتناولها.

6. تخطيط الأدلة:

1.6. تحليل التكرار المشترك

تتمثل المرحلة الأولى في تعيين الأدلة في استكشاف التكرارات المشتركة الأكثر شيوعاً في مجموعة البيانات التي تمت تصنيفها كمؤشر على الارتباط بين المصطلحات. وبعبارة أخرى، يفترض هذا الأمر أنه كلما زاد التكرار المشترك بين مصطلحين كلما زاد الارتباط بين هذين المصطلحين.

تم تضمين المزيد من التطوير وظيفة التمييز بين مدى ظهور المصطلحات المتزامنة المتقاربة عن بعضها البعض. وهذا يعكس الفرضية القائلة بأن المصطلحات المتزامنة التي تحدث بالقرب من بعضها البعض في الجملة أو المشاركة تعتبر أكثر أهمية من تلك المصطلحات التي تحدث متباعدة عن بعضها البعض. من ناحية أخرى، لمراعاة العلاقة النحوية بين المصطلحات المتزامنة في التكرار المشترك، تم استخدام شجرة تبعية؛ وبالتحديد، تم استخدام بنك توركو للتبعية (Turku DependencyTreebank) للتعامل مع اللغة الفنلندية. يظهر في الشكل 3 مثال على استخدام شجرة التبعية في توركو.



الشكل 3 مثال على نتيجة شجرة التبعيات توركو

يسمح لنا استخدام الشجرة الإعرابية بالنقاط العلاقة بين الألفاظ حتى لو كانت متباعدة عن بعضها البعض. على سبيل المثال، الفعل وفاعله، وكذلك مكمّله، يستلزم وجود رابط مباشر من الفعل إلى فاعله ومن الفعل ومكمّله. لذلك، نفترض أنه عندما تكون كلمتان، على سبيل المثال، w_1 و w_2 مرتبطتين ببعضهما البعض مباشرةً في نظر شجرة المُعرب فإن المسافة بين هاتين الكلمتين تساوي واحدًا حتى لو كانتا متباعدتين عن بعضهما البعض، وإلا فإن المسافة تُحسب بعدد الرموز التي تفصل بين w_1 و w_2 . بشكل أكثر رسمية، يمكن صياغة ذلك¹ على النحو التالي:

$$d(w_1, w_2) = \begin{cases} 1 & \text{if } w_1 \text{ and } w_2 \text{ are directly linked in the parser} \\ \text{number of tokens between } w_1 \text{ and } w_2, & \text{otherwise} \end{cases}$$

من منظور تطبيقي، اقتصرنا على خمسة مستويات مختلفة من المسافة: مسافة 1-2 كلمة، أي بجانب بعضها البعض (إما مباشرة أو من خلال بنية شجرة التبعية)، مما يؤدي إلى تأثير قوي على بعضها البعض؛ ومسافة 3-5 كلمات، مما يشير إلى تأثير كبير محتمل على بعضها البعض؛ ومسافة 6-7 كلمات، حيث من المحتمل أن يكون التأثير متوسطًا. وأخيرًا، المستويات من 9-14 كلمة (متباعدة عن بعضها البعض) تشير إلى تأثير بسيط إلى حد ما إلى تأثير منخفض على بعضها البعض، أما المسافة 15+ (أكثر من 15 كلمة متباعدة عن بعضها البعض) فتشير إلى أن التواجد المشترك لا يعتبر له أي تأثير على العلاقة النحوية أو الدلالية أو السببية بين المصطلحين w_1 و w_2 . تسلط الخوارزمية 2 (انظر الشكل 4) الضوء على تطبيق نموذج التكرار المشترك القائم على المسافة. يبرز الشكل 5 التكرارات المشتركة للكلمات مع مراعاة المسافات المختلفة. بينما يقدم الجدول 4 تقريراً عن أهم الأزواج المتشابهة المقابلة للمناقشة العامة والمصطلحات المتعلقة بالسرطان.

الجدول 4. التكرار العام للكلمات المشتركة للمصطلحات المتعلقة بالسرطان وغير المتعلقة بالسرطان

Most repeated words		Most repeated cancer words	
Word	#number	Word	#number
jostain-syystä	1709	maksaa-euroa	251
pari-päivää	1687	hoidot-tietoa	102
pari-viikkoa	1683	hoidot-hoitohen	101
osaa-sanoa	1632	hoidot-tehokkuudesta	101
viikon-päästä	1568	hoidot-varmaa	101
seuraavana-päivänä	1562	äiti-isä	101
päästä-eroon	1474	hoitoa-olemassaolevaa	97
lääkäri-sanoi	1466	hoitoa-tutkimusnäyttöä	97
pitäisi-tehdä	1427	hoidot-Tutkijaryhmän	96
the-and	1411	hoidot-jaottelemisen	96
ottaa-yhteyttä	1186	hoidot-kategorioihin	96
and-the	1168	hoidot-myöntää	96

تكشف قراءة سريعة للنتائج ما يلي:

1. يُظهر التدقيق في الكلمات ذات التكرار الشديد، خارج الخطاب المتعلق بالسرطان، هيمنة الاقتراح التفسيري كما في (jostain-syystä) التي تترجم (لبعض الأسباب) والاقتراحات الزمنية كما في

(pari-viikkoa)، (viikon-pästä) التي تترجم إلى (أسبوعين)، (في أسبوع)، على التوالي. يمكن أيضاً ملاحظة المقترحات الشرطية في (pitäsi-tehda) التي تترجم إلى (يجب أن - تجعل). يتم إبراز المقترحات الشبيهة بالتوصية في الزوج (ottaa - yhteytt)، الذي يترجم إلى (خذ - اتصل).

2. على الرغم من شبه هيمنة اللغة الفنلندية، إلا أن بعض المصطلحات الإنجليزية لا توجد العديد من الصياغات الإنجليزية التي تبرر هيمنة التعبير المحدد (the) وتعبير العطف (and) .

3. هناك تكرار واحد مشترك أكثر ارتباطاً بالبيئة السريرية (lääkäri – sanoi)، وهو ترجمة لـ (طبيب - قال)، مما يشير إلى الاهتمام الكبير لمستخدمي المجتمع بطلب مشورة الطبيب وتوصياته من وجهات نظر الخبرة الأخرى.

4. يشير التدقيق في الخطاب المرتبط بالسرطان إلى هيمنة الجانب المالي حيث يبحث المستخدمون/المريض عن تكاليف العلاج إما بشكل مباشر أو غير مباشر كما في (maksaa-eur). وترتبط المفردات التالية التي تتشارك في التكرار بنوع العلاج وفعاليته (انظر، جميع الأزواج مع كلمة hoidot). يلاحظ المرء أيضاً ظهور القضايا المتعلقة بالأسرة في (äiti-isä)، ترجمة لعلاقة (الأم-الأب).

5. يسلط ما سبق الضوء على أهمية آثار العلاج والتكلفة المالية والمشكلات العائلية كعوامل مهيمنة في خطاب السرطان. بينما يُظهر الخطاب غير المتعلق بالسرطان أهمية المشورة السريرية في المجتمع الصحي في Suomi24 بحثاً عن أقوال الأطباء وتأكيدهم. من ناحية أخرى، تشير هيمنة التعبيرات الشرطية والتوصيات إلى أهمية هياكل الجدول وسيناريوهات سرد القصص حيث يحاول المستخدمون بناء حجج للرد على قلق المريض/المستخدم أو من خلال مشاركة تجربة شخصية.

6. يُظهر تحليل نتائج الشكل 6 التفاوت الكبير في قيم التكرار بين المصطلحات التي تتشارك في التواجد على المسافة 1 (أو أحياناً المسافة 2) وتلك التي تتشارك في التواجد على مسافات أعلى (3-5، 6-8، 9-14، 15+) بمقدار يزيد عن 100! وهذا يشير إلى أن التكرارات المشتركة التي تحدث على مسافة أكثر من 2 غير مهمة إحصائياً عند مقارنتها بتلك التي تحدث على مسافة 1 أو 2.

Algorithm 2 co-occurrence detection and topic from Suomi24

Input: Finnish Cancer Keywords, Suomi24 files, Right Topics, Finnish Stopwords, topics Keywords, database Access Keys

Output: wordCount[topics,wordDistance,isComment,keywords], database

```

1: procedure SUOMI24 ANALYSIS
2:   for every file in Suomi24 files do
3:     items ← file[items]
4:     for every item in items do
5:       topics ← item[topics]
6:       body ← item[body]
7:       for sentence in body and comments do
8:         for word in sentence do
9:           if length of word > 2 and length of the next
word > 2 then
10:            wordCount ← word cooccurrence
11:            if distance between words < 3 then
12:              wordCount[wordDistance] ← "1-2"
13:            else if distance between words < 6 then
14:              wordCount[wordDistance] ← "3-5"
15:            else if distance between words < 9 then
16:              wordCount[wordDistance] ← "6-8"
17:            else if distance between words < 15 then
18:              wordCount[wordDistance] ← "8-14"
19:            else
20:              wordCount[wordDistance] ← "15+"
21:            end if
22:            if word in deathList[] then
23:              increment keywords["death"]
24:            else if word in illnessList[] then
25:              increment keywords["illness"]
26:            else if word in treatmentList[] then
27:              increment keywords["treatment"]
28:            else if word in socialList[] then
29:              increment keywords["social"]
30:            else if word in financialList[] then
31:              increment keywords["financial"]
32:            end if
33:            wordCount[keywords] ← keywords
34:          end if
35:          if inComments then
36:            wordCount[isComment] ← inComments
37:          else if inBody then
38:            wordCount[isComment] ← inTopic
39:          end if
40:          wordCount[topics] ← topics
41:          DATABASE ← wordCount[topics,wordDistance,isComment,keywords]
42:        end for
43:      end for
44:    end for
45:  end for
46: end procedure

```

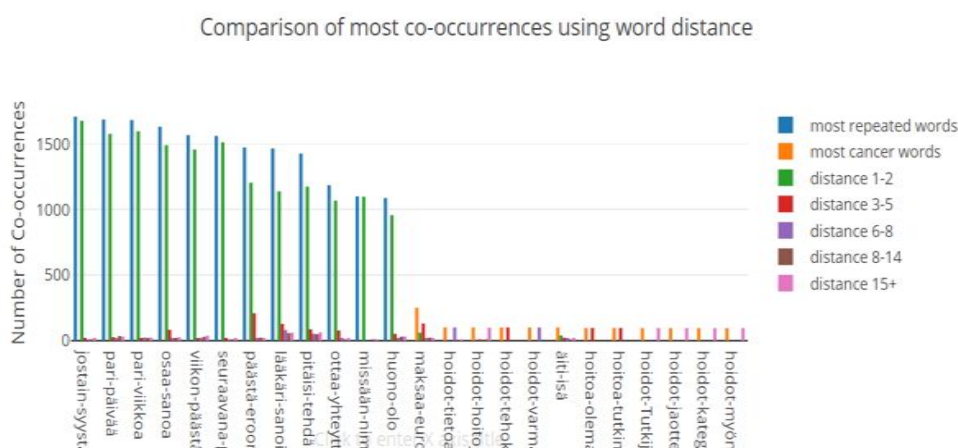
الشكل 4. الكشف عن التكرار المشترك ونمذجة الموضوع

7. تصنيف العوامل النفسية والاجتماعية:

1.7 العقلانية

في هذه الدراسة، تم تصنيف العوامل النفسية والاجتماعية إلى خمس فئات: المالية، والمرضى، والعلاج، والاجتماعية، والوفاة. ويستند الدافع وراء هذا التصنيف إلى التفكير المنطقي ودراسة الأدبيات والتحليل اليدوي للبيانات. في الواقع، من البديهي أن مرضى السرطان يتوقعون الموت كنهاية نهائية سواء على المدى القصير أو المتوسط. بالإضافة إلى ذلك، فإن نوع المرض وفعالية العلاج والتكلفة المرتبطة به (التكلفة السريرية والتكلفة المتعلقة بالأسرة) وكذلك الحياة الاجتماعية للمريض هي بطبيعة الحال ذات أهمية قصوى لأي مريض سرطان. ثانياً، أبلغت الدراسات السابقة عن العوامل السريرية والنفسية الاجتماعية المرتبطة باضطرابات القلق والاكتئاب لدى مرضى سرطان الثدي وغيرهم من مرضى السرطان [15]. وقد حدد المؤلفون ثلاثة أنواع من العوامل النفسية والاجتماعية المرتبطة بها: الدعم الاجتماعي، والعلاقة الأسرية، وحل النزاعات. كان القلق والاكتئاب لدى مريضات سرطان الثدي شائعاً عادةً في بيئات ذات علاقات أسرية ضعيفة، مما ساهم أيضاً في زيادة الألم والإرهاق [7]. وهذا يؤكد أهمية العوامل الاجتماعية كجزء من العوامل النفسية والاجتماعية لمرضى السرطان. أشارت الدراسات في [27] إلى أن المزيد من الدعم الاجتماعي كان مرتبطاً بانخفاض معدل الإصابة بالسرطان والوفيات الناجمة عنه، وفترة أطول من الشفاء من المرض والبقاء على قيد الحياة، مما يثير أهمية عوامل العلاج والمرض والوفاة. وبالمثل، وجد بيترز وآخرون [24] أن عدد الروابط الاجتماعية والعزلة الاجتماعية المتمرسه كانت مرتبطة بالإصابة بالسرطان والوفيات الناجمة عن السرطان. وقد وجد Mofatt وآخرون [19] أن بداية الإصابة بالسرطان وعلاجه ومسار المرض مرتبط بالضغوط المالية بين المرضى، والتي تم تحديدها على أنها حاجة كبيرة لم تتم تلبيتها. وهذا يؤكد أهمية العوامل المالية كجزء من المظلة الشاملة للعوامل النفسية والاجتماعية. ثالثاً، كشف التدقيق اليدوي في العديد من منشورات Suomi24 عن أن نجاح العلاج المحتمل وفشله (الوفاة) يظهران بين مواضيع النقاش الساخنة. وبالمثل، عادةً ما يرتبط المرض والعلاج بالمحادثات المتعلقة بالسرطان، حيث يسعى الأشخاص عادةً إلى التنقيف أو دعم الأقران أو المشورة للتغلب على مرضهم والتحديات المرتبطة به [8]. تُظهر تكاليف الوقت والنفقات من الجيب العبء المالي الكبير الذي يتحمله المرضى عادة، وغالباً ما يكونون بموارد محدودة، ويعتمدون على أفراد الأسرة أو الأصدقاء [22]، مما يعزز دور القضايا المالية كجزء من العوامل النفسية والاجتماعية الرئيسية. رابعاً، إن تحليل تواتر المصطلحات والتكرار المشترك الذي تم إجراؤه في القسم السابق يدعم بلا شك التصنيف أعلاه للعوامل النفسية الاجتماعية. وبالفعل، يكشف التكرار البسيط لمصطلحات خطاب السرطان عن هيمنة الجوانب المالية وآثار العلاج والوفاة والقضايا المتعلقة بالمرض (أو المرضى). بينما يبرز الخطاب غير المتعلق بالسرطان هيمنة العوامل الاجتماعية. في الواقع، كانت العلاقات الأسرية، على سبيل المثال، من بين الكلمات الأكثر تواتراً في الشكل 5. إلى جانب ذلك، فإن وفرة

التعبيرات الجدلية والسرد القصصي، التي تساعد على مشاركة سرد القصص وبناء عبارات مقنعة بطريقة تجيب على مخاوف المريض، تشهد على أهمية الدعم النفسي / الاجتماعي.



الشكل 5. المدرج التكراري للمصطلحات الأكثر تزامنًا ومقارنتها مع المصطلحات المتعلقة بالسرطان مع المسافة

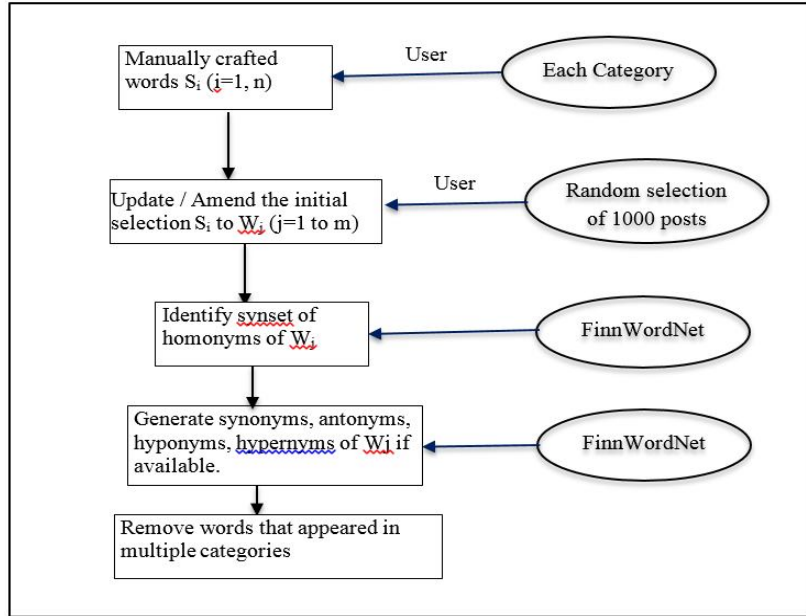
8. التوصيف الأنطولوجي للفئات

استلهاً من التقنيات المستخدمة في بعض أبحاث الكشف عن خطاب الكراهية (انظر، على سبيل المثال، ورقة مراجعة Schmidt و Wiegand^[30])، فإن الفكرة المستخدمة لبناء الأنطولوجيا المتعلقة بكل فئة هي التالية. أولاً، نبدأ بعدد محدود من الكلمات المصاغة يدوياً والمخصصة لكل فئة بناءً على فهمنا للمعنى اللغوي والسياقي لكل فئة. ثانياً، نُدْجِي تحقيقاً يدوياً في مجموعة عشوائية مختارة عشوائياً لحوالي 1000 مشاركة من مجموعة البيانات المجمعة. والهدف من هذا التحقيق هو تنقيح الكلمات التي تمت صياغتها يدوياً لكل فئة من خلال تعديل أو إضافة كلمات إضافية وجدنا أنها ذات صلة بفئة معينة بعد قراءة المنشور (المنشورات)، حيث يحدث غالباً أن العديد من المصطلحات التي تم تحديدها من خلال مراجعة المنشور تكون أقل ارتباطاً من الناحية اللغوية والدلالية بالمعنى الفعلي للفئة. من ناحية أخرى، نظراً لأن العديد من الكلمات المتجانسة ناتجة عن مرحلة التدقيق اليدوي والتدقيق اللاحق العشوائي، فإننا نعيد النظر في كل كلمة متجانسة ونعَي المعنى المقابل (synset) الذي يتطابق بشكل أفضل مع الفئة التي تم تسطيرها. استخدمنا شبكة FinnWordNet، التي قدمتها جامعة هلسنكي كجزء من مشروع FIN-Clam^[6] لتوليد مجموعات المترادفات المرتبطة. يتم إجراء هذه العملية بشكل مستقل من قبل ثلاثة من الناطقين باللغة الفنلندية من أجل القضاء على أي تحيز أو تعارض أو استبعاد بعض الأحداث المثيرة للاهتمام. ثالثاً، باستخدام قاعدة البيانات المعجمية (الفنلندية) WordNet، تم إنشاء نظام آلي لإضافة

مجموعة من المترادفات والمتضادات والمرادفات الإضافية والمرادفات المفردة لكل كلمة من فئة معينة كلما توفرت. إذا تبين أن الكلمة التي تم إنشاؤها حديثاً (مترادف، أو متضاد أو مرادف افتراضي أو مرادف فرعي أو تشعبي) تنتمي إلى أكثر من فئة واحدة، يتم حذفها بشكل منهجي. على الرغم من أن هذا المنطق بديهي ومبتكر في نهجه لتوليد مجموعة كبيرة بشكل معقول من الكلمات ذات الصلة لكل فئة، إلا أن له أيضاً حدوده كما سيتم تسليط الضوء عليه لاحقاً. يظهر الهيكل العام لبناء الأنطولوجيا في الشكل 6.

9. تقييم الفئات في مجموعة البيانات المجمعّة

بعد توليد الأنطولوجيا المرتبطة بكل فئة من فئات العوامل النفسية والاجتماعية، نود تقييم حدوث كل فئة في مجموعة البيانات المجمعّة بشكل عام.



الشكل 6. الهيكل العام لتوليد أنطولوجيا الفئات النفسية والاجتماعية

لهذا الغرض، نظراً لأن أنطولوجيا كل فئة تتكون من مجموعة من المصطلحات التي تم إنشاؤها من خلال عملية الشكل 6، فإن مطابقة سلسلة بسيطة ستتمكننا من تحديد مدى التطابق في مجموعة البيانات الإجمالية. ومع ذلك، يجب إيلاء عناية خاصة لتلك المصطلحات المتجانسة التي يتم تمثيلها من خلال مجموعات مترادفات خاصة بها. في مثل هذه الحالات، يتم اتخاذ خطوة إضافية لمعرفة ما إذا كانت هناك مطابقة للحواس أيضاً أم لا. وبالتالي، هناك حاجة إلى وجود آلية لتحديد المجموعة المترادفة الصحيحة لكلمة مترادفة معينة في المنشور. لهذا الغرض، تم تصميم آلية جديدة تستخدم نهج التشابه الدلالي. بشكل أكثر تحديداً، بالنظر إلى المنشور P، الذي تم ترميزه على أنه

$$P = \langle T_1, T_2, \dots, T_m \rangle$$

لنفترض، على سبيل المثال، دون فقدان العمومية، أن T_2 هي الكلمة الهدف التي يسعى المرء إلى تحديد مجموعة synset المرتبطة بها.

1. هناك نهج أعمى لا يستكشف العلاقة النحوية الكاملة للجملة التي تحتها خط والتي تحتوي على T_2 ولكنه جذاب من الناحية الحسابية يتكون من (i) حساب التشابه الدلالي لـ T_2 مع أول مصطلح يقع بجواره والذي يوجد له مثل هذا التشابه الدلالي باستخدام، على سبيل المثال، تشابه شبكة كلمات Wu و Palmer (Sim_{WP}) [36]، ثم (ii) استرجاع المجموعة المتزامنة لـ T_2 التي تنتج درجة التشابه هذه. بعبارة أخرى، يتلخص ذلك فيما يلي:

*If Sim_{WP}(T₁, T₂) exists then retrieves synset (T₂)
 else If Sim_{WP}(T₂, T₃) exists then retrieves synset (T₂)
 else If Sim_{WP}(T₂, T₄) exists then retrieves synset (T₂)
 Etc.*

إن الحدس الكامن وراء هذا المنطق هو أن درجة التشابه الدلالي لشبكة الكلمات بين كلمتين T_i و T_j تنتج دائماً أقصى درجة يمكن تحقيقها عند النظر في جميع مجموعات متلازمات T_i و T_j . وبالتالي، تتوافق مجموعة المترادفات التي تم إنشاؤها مع المعنى الأكثر توافقاً مع الكلمة التالية في المنشور. وهذا يجسد إلى حد كبير المعنى الصحيح للكلمة الهدف في الجملة التي تحتها خط. يتم وصف التنفيذ الخوارزمي للخوارزمية السابقة في الخوارزمية 3.

Algorithm 3: Synset identification of homonym target word in the post

Input: 1. homonym word W_k

2. Post sequence: $W_1, W_2, \dots, W_k, W_{k+1}, \dots, W_n$

Output: Synset W_k^t

$t=k$

0: If Sim_{WP}(W_{k-1}, W_t) exists Return synset W_k^t of W_k

Else If Sim_{WP}(W_t, W_{k+1}) exists Return synset W_k^t of W_k

Else $k = k + 1$ GOTO 0

END

بعد ذلك، بمجرد الانتهاء من عملية مطابقة الكلمات المتجانسة مع المنشور في المنشور، يتحول التقييم الكلي إلى عملية مطابقة سلسلة بسيطة كنسبة من حدوث كل عامل في مجموعة البيانات المجمعة بأكملها. يتم تنفيذ ذلك باستخدام منهجية بسيطة قائمة على مطابقة الكلمات. بشكل أكثر رسمية، بالنظر

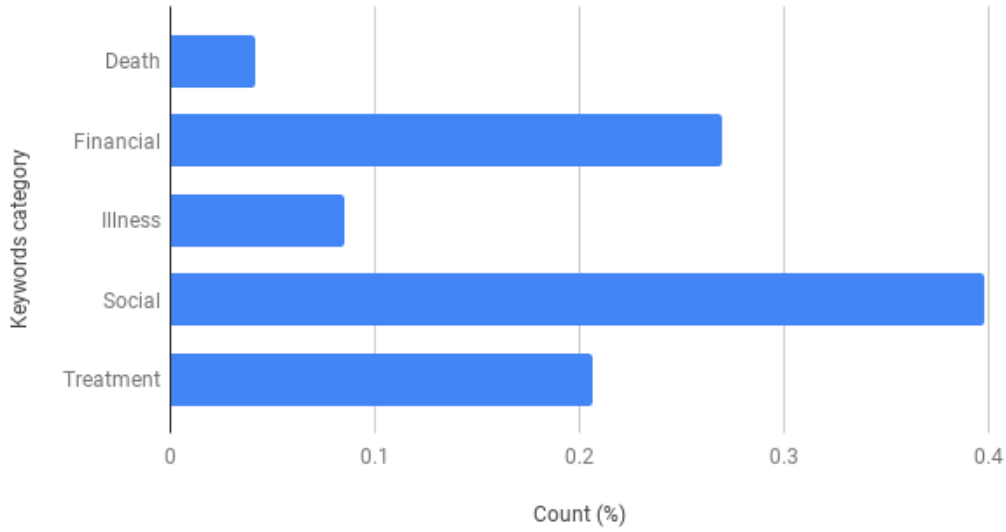
إلى الفئة i^{th} أمن العامل النفسي الاجتماعي (المرض، والعلاج، والاجتماعي، والمالي، والوفاة)، يتم حساب النسبة المرتبطة بها p_i على النحو التالي:

$$p_i = \frac{1}{c} \sum_j |\lambda_{C_i \cap Post_j}|, \quad (2)$$

$$\lambda_{C_i \cap Post_j} = \begin{cases} 1 & \text{if } C_i \cap Post_j \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

حيث يشير c إلى عامل التطبيق، $Post_j$ إلى مجموعة الرموز المرتبطة بالمنشور i^{th} أمن مجموعة البيانات، بينما يشير C_i إلى جميع الصياغات (الرموز) التي يمكن ربطها بالفئة النفسية والاجتماعية i^{th} . يتم إجراء المجموع في التعبير أعلاه على جميع المشاركات الواردة في مجموعة البيانات. يوضح الشكل 7 قوة كل عامل من العوامل النفسية والاجتماعية الخمسة.

Count (%) vs. Keywords category

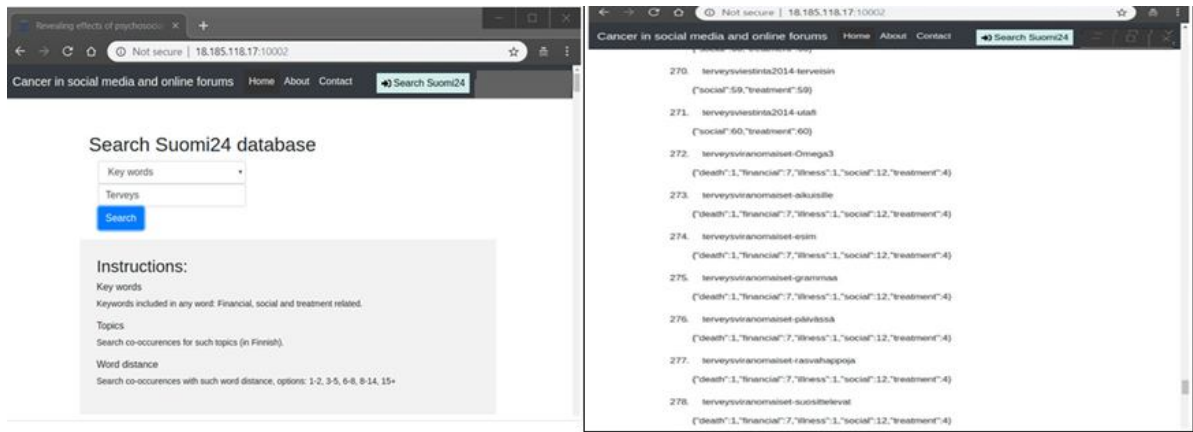


الشكل 7. النسبة المئوية لتصنيف العوامل النفسية والاجتماعية

تُظهر النتائج هيمنة العوامل الاجتماعية تليها المشاكل المالية والمشاكل المتعلقة بالعلاج. ومن البديهي أن مرضى السرطان يعانون من العديد من المخاوف مثل كيفية تعامل أزواجهم أو أطفالهم مع المرض بعد إصابتهم بالسرطان، في حين أن العبء المالي قد ينشأ بسبب الخوف من فقدان الراتب بسبب المرض، أو احتمال خفض الراتب. يرتبط هذا أيضاً بشكل إيجابي مع نتائج أبحاث أخرى. على سبيل المثال، يؤكد كل من McFarland و Holland^[16] و Falisi وآخرون^[7] على ميل مرضى السرطان إلى بناء حياة اجتماعية تتغلب على الآثار النفسية للعلاج والمرض، في حين ينشأ العبء المالي بشكل طبيعي كتكلفة ضمنية إما للعلاج أو للتكاليف المتعلقة بالأسرة.

10. تنفيذ النظام:

لتمكين حل مرّن يستكشف مجموعة بيانات Suomi24، بالإضافة إلى توفير تجربة مستخدم كافية، قمنا بتنفيذ تطبيق ويب مدعوم بقاعدة بيانات خلفية. وكمُنصة، تم استخدام خادم ApacheHTTP Server كمنصة، بينما تم اختيار Firebase لقاعدة البيانات الخلفية. تمت كتابة التطبيق بشكل أساسي بلغة JavaScript مع مكتبة React. يتيح استخدام تطبيق الويب سهولة الوصول، ويمكن كلاً من الأطباء والباحثين الأكاديميين من اختبار فرضيات بحثية مختلفة واكتساب رؤى مهمة فيما يتعلق بسلوك مرضى السرطان كما هو موضح في منتدى مناقشة Suomi24. يُظهر الشكل 10 واجهة مستخدم لتطبيق الويب مع استعلام، حيث "Terveys"، والتي تعني "الصحة"، هي أحد الموضوعات الفرعية في مناقشات Suomi24. وتظهر نتيجة الاستعلام في نفس الرسم البياني، حيث يتم عرض التكرارات المشتركة وجميع الموضوعات الفرعية التي حدثت في المناقشات المكتشفة، مع إحصاء عدد مرات حدوثها.



الشكل 8. واجهة البحث من تطبيق الويب.

إلى اليسار: واجهة البحث من تطبيق الويب. إلى اليمين: نتيجة البحث لمثال الاستعلام

1.10. تصور الآثار في حال thread والعوامل النفسية والاجتماعية

إلى جانب ذلك، تم تصميم واجهة أيضاً من أجل تصور تأثيرات البيانات الوصفية الأخرى لمجموعة بيانات Suomi24 لكل مصطلح (متكرر بشكل كبير) من المصطلحات المشتركة (المتداولة بشكل متكرر). وهذا يشمل، على سبيل المثال، سلاسل المناقشة (المواضيع) التي تتكرر فيها هذه المصطلحات بشكل مشترك، وتباين المسافة بين الكلمات المتشابهة بين المصطلحات المتشابهة، والعوامل الاجتماعية النفسية والاجتماعية التي تم تحديدها عبر مختلف المنشورات التي تتكرر فيها المصطلحات. يوضح الشكل 11 لقطة من قاعدة البيانات النارية مع مثال واحد. وتتضمن الوظائف الأخرى ما إذا كانت

المصطلحات المتداخلة قد حدثت في عنوان الموضوع أو المواضيع الفرعية أو في نص المنشور فقط، والفترة الزمنية وما إذا كان اسم المستخدم الخاص بالمؤلف متاحاً.

11. مناقشات

على الرغم من جاذبية التحليل المذكور أعلاه بسبب استقلاليته الكاملة وقدرته على استيعاب المعرفة الخبيرة من حيث الوصف الأنطولوجي لمختلف المفاهيم المتعلقة بالصحة التي تم بحثها خلال هذه الدراسة، إلا أنه لا يخلو من قيود متأصلة.

أولاً، كما هو الحال في معظم تقنيات تحليل البيانات، تتأثر النتيجة بالقيود الجوهرية لتقنيات التتقيب عن النصوص. ويشمل ذلك ⁽¹⁾ تأثيرات تطبيع الكلمات، والتي قد تلغي بعض العناصر المهمة في تراكيب الجمل وتجعل مخرجات المحلل التحليلي غير صحيحة؛ ⁽²⁾ تأثيرات الكلمات المركبة، والتي لا يحددها المحلل التحليلي أحياناً على هذا النحو، مما يؤثر بشكل تافه على أداء مهمة استرجاع المعلومات؛ ⁽³⁾ وجود تعبيرات متعددة اللغات، حيث توجد في كثير من الأحيان تعبيرات إنجليزية وسويدية في مجموعة بيانات Suomi24، مما يجعل مخرجات المحلل التحليلي ضعيفة جداً في بعض الأحيان.

ثانياً، يمثل التعامل مع اللغة الفنلندية صعوبة إضافية بمعنى أن الأدوات الموجودة، بما في ذلك أدوات التحليل وأدوات تطبيع الكلمات وشبكة FinnWordNet الفنلندية لا تزال غير مستقرة مقارنةً بأدوات اللغة الإنجليزية. يؤثر هذا بشكل بديهي على دقة النتائج أيضاً. ولذلك، فإن ناتج تفكيك معنى الكلمات للكلمات المتجانسة، على سبيل المثال، يخضع لعدم استقرار شبكة FinnWordNet وحساب درجة التشابه الدلالي اللاحق. ومع ذلك، تجدر الإشارة إلى أن الترادف المتجانس يؤثر إلى حد كبير على جزء صغير فقط من كلمات أنطولوجيا الفئات، مما يحد من نطاق هذا القيد على نتائج التقييم.

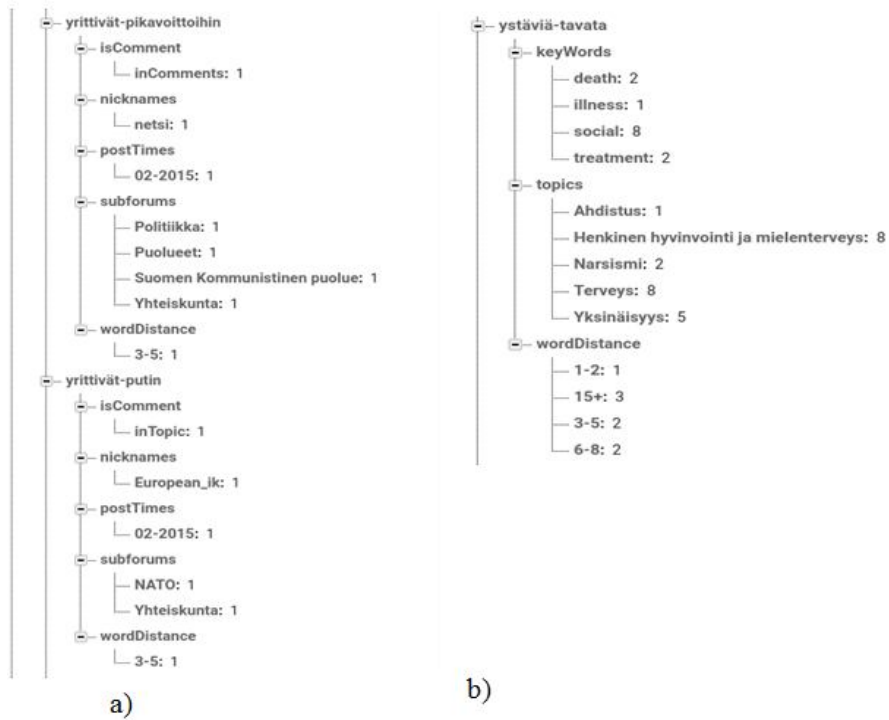
ثالثاً، يخضع استخدام الصياغة اليدوية الأولية للتصميم الأولي للأنطولوجيا التصميمية الأولية المرتبطة بكل فئة بشكل بديهي للحكم الذاتي للمؤلفين المسؤولين عن التسمية اليدوية وفقاً لفهمهم لنطاق كل فئة. إلى جانب ذلك، فإن الغموض المتأصل المرتبط بالتفسير اللغوي البشري يجعل من الصعب جداً تجنب الصياغة المتداخلة التي يمكن أن تنتمي بطبيعة الحال إلى أكثر من فئة واحدة.

رابعاً، إن اللجوء إلى اختيار عشوائي للمشاركات لإضافة/تعديل الصياغة الأولية يخضع لطبيعة العشوائية التي تتسم بها تلك المشاركات. على الرغم من أن اللجوء إلى القراءة اليدوية للمشاركات يمكننا من اكتشاف تعابير اصطلاحية جديدة استخدمها المستخدمون كان من المستحيل استنتاجها باستخدام تفسيرات لغوية بسيطة لنطاق الفئات فقط، إلا أن مدى هذه الصياغة الاصطلاحية بعيد عن أن يكون

شاملاً. وبالفعل، قد يتم التغاضي عن العديد من الصياغات المهمة الصادرة عن المصطلحات المحلية التي تصف جانباً من جوانب المالية أو العوامل الاجتماعية، على سبيل المثال. وبالتالي، لا يُستبعد أن يؤدي التدقيق اليدوي الإضافي للمشاركات إلى مزيد من التوسع في خطاب أنطولوجيا الفئات. ومع ذلك، فإننا نعترف أنه على عكس التدوين المصغر، فإن مدى هذه المصطلحات ليس وفيراً في Suomi24. ويتفق هذا أيضاً مع بعض الدراسات الأخرى ذات الصلة التي شملت مجموعة Suomi24 كجزء من مشروع "عقول المواطنين الفنلنديين" التابع لأكاديمية فنلندا^[17].

خامساً، تكشف نتيجة المرحلة الأولى والثانية في بناء الأنطولوجيا عن وجود عدة صياغات تنتمي إلى نفس الكلمة الجذرية، ولكنها تنتمي إلى فئات مختلفة. وبناءً على ذلك، فإن مثل هذه الصياغات من شأنها أن تستحث نفس الترادفات والمرادفات الزائدة، مما يؤدي إلى إلغاء هذه الكلمات كجزء من البناء الأنطولوجي الكلي. وبالمثل، وبغض النظر عما إذا كانت الكلمات المذكورة أعلاه جزءاً من نفس تصنيف الكلمة أم لا، فإن استخدام علاقة الترادف والترادف المفرط قد يؤدي أيضاً إلى كلمات تم تعيينها بالفعل إلى فئات أخرى.

سادساً، يمكن أيضاً التشكيك في فكرة الحذف المنهجي للكلمات التي تنتمي إلى فئات متميزة كجزء من منهجية التصميم. في الواقع، على الرغم من أن هذا الأمر يتوقع أن يقلل من التداخل المحتمل بين هذه الفئات، والذي بدوره يزيد من إمكانية تمييزها، إلا أنه يمكن أن يؤدي إلى التقليل من شأن تكرارها. وبعبارة فضفاضة، فإن تأثير هذا التقليل من أهمية هذا التقليل ليس مهماً لأن الهدف الأساسي هو بالأحرى فهم مدى الاختلافات الموجودة بين تكرارات هذه الفئات في مجموعة المواد بدلاً من التقدير الدقيق لحدوث كل فئة على حدة.



الشكل 9. عينة من المخرجات على قاعدة البيانات (Firebase)

الخلاصة:

توفر هذه الورقة البحثية منصة مفتوحة لتحليل البيانات الصحية لـ Suomi24 من أجل تحديد العوامل النفسية والاجتماعية المرتبطة بمرضى السرطان. تستفيد المنهجية من المعالجة المبتكرة للغة الطبيعية المبتكرة التي تستدعي بناء أنطولوجيا جديدة لكل من جمع البيانات وتصنيفها بالإضافة إلى تحليل التكرار المشترك متعدد المستويات بطريقة تحدد قوة الكلمات المتزامنة. من ناحية أخرى، تم طرح تصنيف للعوامل النفسية والاجتماعية مستوحى من تحليل العوامل النفسية والاجتماعية في الدراسات الطبية ذات الصلة، وتم طرح تصنيف للعوامل النفسية والاجتماعية والتحقيق فيها. وتعزز النتائج بعض النتائج السابقة في الدراسات السريرية المتعلقة بهيمنة العوامل الاجتماعية. تم طرح منصة مفتوحة لمساعدة الباحثين والأطباء السريريين على حد سواء. يمهّد العمل الطريق لمزيد من التحقيق في البيانات النصية الطبية من أجل ربط النتائج بسجلات المرضى كما هي مسجلة في السجل الطبي الإلكتروني. إلى جانب ذلك، يسلط التحليل الضوء أيضاً على التحدي المتمثل في استيعاب أدوات البرمجة اللغوية العصبية القياسية للغة الفنلندية حيث يلزم بذل المزيد من العمل من أجل مطابقة أداء الأدوات في النص الإنجليزي. على وجه الخصوص، يمكن استخدام مفهوم الحجة النصية وتلخيص النص، استمراراً لعملنا في [20، 21]، لضمان بعض الاتساق والتماسك المتأصل في المخرجات النصية.

Référence :

1. Adamic L.A., and Adar E. (2003). Friends and neighbors on the Web. *Social Networks*, 25(3), 211-230
2. Adar E., and Adamic L.A. (2006). Tracking Information Epidemics in Blogspace. In: *Proceedings of Web Intelligence*, Compiegne.
3. Akay A., Dragomir A., Erlandsson B.E., and Akay M. (2016). Assessing antidepressants using intelligent data monitoring and mining of online fora, *IEEE J. Biomed. Health Informat.*, 20(4), 977-986.
4. Bao S., Xu S., Zhang L., Yan R., Su Z., Han D., Yu Y. (2012). Mining social emotions from affective text, *IEEE Trans. Knowl. Data Eng.*, 24(9), 1658-1670.
5. Falisi A.L., et al. (2017). Social media for breast cancer survivors: a literature review. *J Cancer Surviv*, vol.11, 808–821.
6. FinnWordNet (2019), <http://www.ling.helsinki.fi/en/lt/research/finnwordnet/>
7. Garssen B. (2002). Psycho-oncology and cancer: linking psychosocial factors with cancer development. *Annals of Oncology*, 13(4), 171-175.
8. Helgeson V.S., Cohen S., Schulz R., Yasko J. (1990). Education and peer discussion group interventions and adjustment to breast cancer. *Archives of General Psychiatry*, 56(4), 340-347.
9. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000). Self organization of a massive text document collection. *Transactions on Neural Networks. Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3), 574-585.
10. Kummervold P. E., Gammon D., Bergvik S., Johnsen J.K., Hasvold T., Rosenvinge J.H. (2002). Social support in a wired world: use of online mental health forums in Norway. *Nord. J Psychiatry*, 56(1), 59-65.
11. Lagus K. M., Pantzar M., Ruckenstein M., and Ylisiurua M.. (2016), *Suomi24 – Muodonantoa Aineistolle*. Technical Report, University of Helsinki.
12. Lagus K.H., Pantzar M., Ruckenstein M.S., Ylisiurua M.J. (2016). Suomi24 - Muodonantoa aineistolle (Suomi24 - Giving shape to the data set), in: *Proceedings of the Valtiotieteellisen tiekunnan julkaisuja Workshop*, Helsinki.
13. Lagus K., Ruckenstein M., Juvonen A., and Rajani C. (2018). Medicine Radar – A tool for exploring online health discussions, *Proceedings of the Digital Humanities in the Nordic Countries, CEUR Workshop*, Vol. 2084, 460-468.
14. Lento T., Welser H.T., Gu L., and Smith M. (2006). The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System, In: *Proceedings of the third annual workshop on the weblogging Ecosystem: Aggregation, Analysis, and Dynamics*. Edinburgh, Scotland.
15. Lueboonthavatchai P. (2017). Prevalence and psychosocial factors of anxiety and depression in breast cancer patients. *Journal-Medical Association of Thailand*, 90(10), 2164.
16. McFarland D.C., Holland J.C. (2016). The management of psychosocial issues in oncology. *Clinical Advances in Hematology & Oncology*, vol.14, 999–1009.
17. Mindscapes (2019). Citizen Mindscapes Project. <https://blogs.helsinki.fi/citizenmindscapes/>
18. Mitchell A.J., Chan M., Bhatti H., Halton M., Grassi L., Johansen C., Meader N. (2011). Prevalence of depression, anxiety, and adjustment disorder in oncological,

- haematological, and palliative-care settings: A meta-analysis of 94 interview-based studies. *The Lancet Oncology*, 12(2), 160–174.
19. Moffatt S., Noble E., White M. (2017). Addressing the Financial Consequences of Cancer: Qualitative Evaluation of a Welfare Rights Advice Service. *PLoS ONE* 7(8): e42979. <https://doi.org/10.1371/journal.pone.0042979>
 20. Mohamed M., and Oussalah M. (2020), SRL-ESA-TextSum. A Text Summarization Approach Based on Semantic Role Labeling and Explicit Semantic Analysis, *Information Processing and Management*, <https://doi.org/10.1016/j.ipm.2019.04.003>
 21. Mohamed M., and Oussalah M. (2019). A Hybrid Approach for Paraphrase Identification Based on Knowledge-enriched Semantic Heuristics, *Language Resources and Evaluation*, <https://doi.org/10.1007/s10579-019-09466-4>
 22. Northouse L., Williams A.L., Given B., McCorkle R. (2012). Psychosocial care for family caregivers of patients with cancer. *Journal of Clinical Oncology*, 30(11), 1227-1234.
 23. Nguyen T., Phung D., Dao B., Venkatesh S., Berk M. (2014), Affective and content analysis of online depression communities, *IEEE Trans. Affective Comput.*, 5(3), 217-226.
 24. Peters M.E, Goedendorp M.M., Verhagen S.A., Van-Der-Graaf W.T., Bleijenbergh G. (2014). Exploring the contribution of psychosocial factors to fatigue in patients with advanced incurable cancer. *Psycho-Oncology*, 23(7), 773-779.
 25. Ren F., Kang X., Quan C. (2016). Examining accumulated emotional traits in suicide blogs with an emotion topic model, *IEEE Journal of Biomedical and Health Informatics*, 20(5), 1384-1396.
 26. Rodrigues R.G., Das Dores R.M., Camilo-Junior C.G., Rosa T.C. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks, *International Journal of Medical Informatics* 85(1), 80-95.
 27. Reynolds P., and Kaplan G.A. (1990). Social connections and risk for cancer: prospective evidence from the Alameda County Study. *Behavioral Medicine*, 16(3), 101-110.
 28. Saha B., Nguyen T., Phung D., and Venkatesh S. (2016). A framework for classifying online mental health related communities with an interest in depression, *IEEE J. Biomed. Health Informat.*, 20(4), 1008-1015.
 29. Salvetti F., and Nicolov N. (2006). Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach, In: *Proceedings of HLT-NAACL 2006: Human Language Technology Conference*. New York City, USA.
 30. Schmidt A., and Wiegand M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain, 1–10
 31. Sheldon C.R., Mermelstein R., Kamarck T., and Hoberman H. R. (1985). Measuring the Functional Components of Social Support. In *Social Support: Theory, Research, and Applications*, edited by I.G. Sarason and B.R. Sarason. Boston: Martinus Nijhoff, 73-94.
 32. Suler J. (2004), The online disinhibition effect. *Cyberpsychol Behav*, 7(3), 321–326.
 33. Suomi24 (2019). Suomi24 dataset corpus, available at <http://metashare.csc.fi/repository/download/b4db73da85ce11e4912c005056be118ea699d93902fa49d69b0f4d1e692dd5f1> Archived by WebCite at <http://www.webcitation.org/6oeEAXcAo>.

34. Tai C.H., Tan Z.H., and Chang Y.S. (2016). Systematical approach for detecting the intention and intensity of feelings on social network, *IEEE J. Biomed. Health Informat.*, 20(4), 987-995.
35. Thelwall M. (2004), *Link Analysis: An Information Science Approach*. Academic Press
36. Wu Z., and Palmer M. (1994). Verbs semantics and lexical selection, In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 133-138.
37. Yi J. (2005), Detecting buzz from time-sequenced document streams. In *Proceedings of the IEEE conference on eTechnology, e-Commerce and e-Service*, (EEE '05), 347-352.
38. Zhang L., Hall M., Bastola D. (2018). Utilizing Twitter data for analysis of chemotherapy", *International Journal of Medical Informatics*, 120, 92-100.

-
1. <http://members.unine.ch/jacques.savoy/clef/>
 2. [http://turkunlp.github.io/Finnish-dep-parser /](http://turkunlp.github.io/Finnish-dep-parser/)
 3. <https://github.com/mpsilfve/FinnPos>