การวิเคราะห์ข้อความภาษาไทยเกี่ยวกับการตั้งครรภ์ด้วยวิธีการสร้างแบบจำลองหัวข้อ

ปิยวรรณ ทองพลอย[1*], โสภณ มงคลลักษมี[2], ศิริสรรพ เหล่าหะเกียรติ, รัตน์ชัยนันท์ ธรรมสุจริต[3]

บทคัดย่อ

ปัจจุบันมีผู้ใช้งานอินเตอร์เน็ตจำนวนมากเข้าปรึกษาปัญหากับแพทย์ผ่านชุมชนออนไลน์ (Community question answering: CQA) ซึ่งเป็นงานที่หนักสำหรับแพทย์ที่ต้องตอบคำถามให้ทัน ซึ่งปัจจุบันมีการนำระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) มาประยุกต์ใช้ในการให้ข้อมูล ถามตอบปัญหา แต่การพัฒนาระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) นั้นมีข้อจำกัด เช่น มีราคาแพง และเป็นงานที่ท้าทายในการที่จะทำให้คอมพิวเตอร์เข้าใจถึงภาษามนุษย์จากข้อความในเอกสารที่เป็นภาษาไทย งานวิจัยนี้จึงมีวัตถุประสงค์ในการวิเคราะห์หาคำที่มีนัยสำคัญและการจำแนกหัวข้อในข้อความที่มีความคล้ายคลึงกัน เพื่อสร้างแบบจำลองที่สามารถนำไปพัฒนาระบบโต้ตอบอัตโนมัติ (Chatbot) ให้สามารถโต้ตอบกับผู้ใช้งานได้ตรงประเด็นมากขึ้น โดยใช้วิธีการสร้างแบบจำลองหัวข้อ (Topic modeling) และการจำแนกกลุ่มข้อความ (Clustering) โดยงานวิจัยนี้ได้ประยุกต์ใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ในการวิเคราะห์หาคำที่มีนัยสำคัญและจำแนกหัวข้อในข้อความเกี่ยวกับเรื่องของการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิด ในส่วนของการประเมินประสิทธิภาพ ได้ใช้การวัดผลแบบ extrinsic evaluation โดยใช้เทคนิค K-means ในการจัดกลุ่มหัวข้อและประเมินประสิทธิภาพการจำแนกกลุ่มด้วย Silhouette Coefficient

คำสำคัญ : แบบจำลองหัวข้อ, การจำแนกกลุ่ม, การจัดสรรดีรีเคลแฝง

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[3] คณะแพทยศาสตร์ โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล กรุงเทพฯ 10400

* Corresponding author:Tel.: 083-4442476 E-mail address: piyawan.tho@g.swu.ac.th

THE ANALYSIS OF THAI LANGUAGE ON PREGNANCY PROBLEMS DOMAIN: USING LATENT DIRICHLET ALLOCATION FOR TOPIC MODELING

Piyawan Thongploy[1*], Sophon Mongkolluksame[2], Sirisup Laohakiat, Ratchainant Thammasudjarit[3]

Abstract

Nowadays, there are many messages used to discuss problems with doctors through online communities. It is a difficult task for doctors to answer questions in a timely manner. Currently, chatbot systems are applied to provide information, ask and answer questions, but the development of chatbot systems is limited. For example, it is expensive and a challenging task to make computers understand human language from text in documents. The objective of this research is to analyze significant words and topics in the text, to create a model that can be used to develop an automated response system (Chatbot), to be able to interact with users more relevantly by using topic modeling and clustering methods. The technique used in this research is Latent Dirichlet Allocation in the Analysis of Significant Words and Topics in Pregnancy sexual relations and birth control Texts. Also, as a part of the performance assessment, the K-means technique was used for clustering topics and to assess the cluster efficiency using Silhouette Coefficient.

Keywords : Latent Dirichlet Allocation, Topic modeling, Clustering

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[3] Faculty of Medicine (Ramathibodi), Mahidol University, Bangkok, 10400, Thailand

* Corresponding author:Tel.: 083-4442476 E-mail address: piyawan.tho@g.swu.ac.th

บทนำ

มีผู้ใช้งานอินเตอร์เน็ตจำนวนมากเข้าปรึกษาปัญหาสุขภาพกับแพทย์ผ่านชุมชนออนไลน์ (Community question answering: CQA) ในการปรึกษาเกี่ยวกับเรื่องการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิด การขอคำปรึกษาผ่านชุมชนออนไลน์นั้นเป็นช่องทางที่สะดวกต่อการใช้งาน เพราะสามารถปิดบังตัวตนและสามารถเข้าปรึกษาแพทย์ได้ตลอดเวลา แต่ผู้ใช้งานต้องรอการตอบกลับจากแพทย์ เพราะแพทย์ไม่สามารถให้คำปรึกษาได้ทันทีเมื่อผู้ใช้งานขอคำปรึกษา (Availability) และหากมีปริมาณคำปรึกษามาก แพทย์ไม่สามารถ ให้คำปรึกษาได้ทัน (Scalability) ซึ่งจากข้อจำกัดเหล่านี้ ปัจจุบันได้มีการนำระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) มาประยุกต์ใช้ในการให้ข้อมูล ถามตอบปัญหา การวิเคราะห์ข้อมูล หรือการเก็บข้อมูลจากผู้ใช้งาน ในด้านธุรกิจต่างๆ ทั้งในด้านการบริการลูกค้า ด้านการศึกษา ด้านการบริหารทรัพยากรบุคคล ด้านการเงิน และด้านสุขภาพ [1, 3, 9]

ระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) ถูกพัฒนาขึ้นมาให้มีบทบาทในการตอบกลับการสนทนาผ่านตัวอักษรแบบอัตโนมัติผ่านแพลตฟอร์มต่างๆ ซึ่งกลายมาเป็นที่นิยมในกลุ่มของธุรกิจหลากหลายด้าน โดยระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) นั้นมี 2 ประเภทหลัก คือ แชทบอทที่มีการกำหนดการถามตอบชัดเจน (Rule based Chatbot) และ แชทบอทเพื่อการสนทนา (Conversation Chatbot) ที่มีการนำเทคโนโลยีปัญญาประดิษฐ์ หรือ Machine learning [13] เข้ามาประยุกต์ใช้ให้ระบบสามารถสนทนามีการตอบกลับคู่สนทนาได้เสมือนมนุษย์จริงและสามารถตอบกลับได้ทันที แต่การพัฒนาระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) นั้นมีข้อจำกัด เช่น มีราคาแพง ไม่เข้าใจภาษาท้องถิ่นของผู้ใช้งาน การสะกดคำผิดหรือการใช้คำแสลงบางคำอาจทำให้การตอบไม่ถูกต้อง [10] ซึ่งการประมวลผลภาษาธรรมชาติ หรือ Natural language processing (NLP) ยังเป็นงานที่ท้าทายในการที่จะทำให้คอมพิวเตอร์เข้าใจถึงภาษามนุษย์จากข้อความในเอกสาร โดยวิธีการสร้างแบบจำลองหัวข้อ (Topic modeling) โดยเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation หรือ LDA) เป็นเทคนิคที่นิยมอย่างแพร่หลายในการประมวลผลภาษาธรรมชาติ เพื่อค้นหาหัวข้อและคำที่มีนัยสำคัญที่ถูกซ่อนในเอกสาร [2, 6]

จากการศึกษางานวิจัยมีการนำ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation หรือ LDA) ในการประยุกต์ใช้กับระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) ได้นำมาจำแนกข้อความจากการรับข้อมูลจากผู้ใช้งานหาคำที่มีนัยสำคัญ [11] และ การแปลงข้อความเป็นเวกเตอร์ [8] และได้ถูกมาพัฒนาในหลายๆ ด้าน ตัวอย่างเช่น ในด้านการดูแลสุขภาพ (Healthcare) ได้นำมาประยุกต์ใช้ในการ วิเคราะห์ฟอรัมการตั้งครรภ์ออนไลน์เพื่อต้องการเข้าใจผู้หญิงตั้งครรภ์ให้มากขึ้น [14] ในงานด้านการจำแนกเอกสารเพื่อพัฒนาระบบการจัดประเภทเอกสารวิจัย [7] การนำมาหาข้อเท็จจริงในข่าว โดยการนำมาเปรียบเทียบข้อเท็จจริงระหว่างองค์การอนามัยโลก (WHO) และจากข่าวในช่วงเวลานั้นๆ [5] สรุปเอกสารข่าวออนไลน์หลายฉบับเพื่อให้ผู้อ่านเข้าใจเจตตาของเอกสารข่าวออนไลน์ [12] มีการประยุกต์ใช้ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) กับในหลายๆ โดเมน ซึ่งนำมาใช้กับการวิเคราะห์ข้อความจำนวนมาก

ในงานวิจัยนี้จึงได้ศึกษาการสร้างแบบจำลองหัวข้อ (Topic Modeling) โดยการใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) เพื่อจัดกลุ่มข้อความที่มีลักษณะคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน เพื่อพัฒนาแบบจำลองที่สามารถนำไปพัฒนาระบบโต้ตอบอัตโนมัติ (Chatbot) ให้สามารถโต้ตอบกับผู้ใช้งานได้ตรงประเด็นมากขึ้น ในเรื่องของการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิดในข้อความภาษาไทย โดยมีการวัดประสิทธิภาพของแบบจำลองหัวข้อวัดประสิทธิภาพได้ใช้การวัดผลแบบ Extrinsic evaluation ด้วยการจำแนกกลุ่ม (Clustering) ของข้อความโดยใช้ K-mean โดยประเมินประสิทธิภาพด้วย Silhouette Coefficient

งานวิจัยที่เกี่ยวข้อง

ในส่วนนี้จะมุ่งเน้นการอธิบายการประยุกต์ใช้ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ในการวิเคราะห์หาคำที่มีนัยสำคัญและหัวข้อที่ถูกซ่อนอยู่ในเอกสาร ซึ่งการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) (3) เป็นแบบจำลองในการสร้างคลังข้อมูลด้วยความน่าจะเป็น แนวคิดพื้นฐานคือเอกสารประกอบด้วยการผสมของหัวข้อแฝงแบบสุ่ม ซึ่งแต่ละหัวข้อมีลักษณะเฉพาะที่ถูกแจกแจงด้วยคำต่างๆ โดยจากการศึกษางานวิจัยมีงานวิจัยที่เกี่ยวข้อง ดังนี้

งานวิจัยของ H. Sheikha [5] ได้เสนอการทำเหมืองข้อมูลเกี่ยวกับ covid-19 โดยมีการใช้อัลกอริทึม Latent Semantic Analysis (LSA) และ Latent Dirichlet Allocation (LDA) ในการลดมิติของข้อมูล ซึ่งจะถูกจัดกลุ่มโดยใช้อัลกอริทึม HDBSCAN และ K-Means โดยผลสรุปพบว่า จากรายงานไม่มีความสัมพันธ์ของข้อมูลโซเชียลมีเดียกับข้อมูล WHO โดยตรง แต่มีความสัมพันธ์กับการพาดหัวข่าวโดยการเปรียบเทียบคำที่ใช้บ่อยของกลุ่มคำในวันเดียวกัน โดย HDBSCAN และ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ให้ผลลัพธ์ดีที่สุด

Kim, S.-W., & Gil, J.-M. [7] ได้เสนอระบบการจำแนกเอกสารการวิจัยที่มีความคล้ายคลึงกัน โดยใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ในการดึงคำที่มีนัยสำคัญจากบทคัดย่อ (Abstract) ของแต่ละบทความและตามหัวข้อ จากนั้นใช้อัลกอริทึมการจัดกลุ่ม โดยใช้ K-Means เพื่อใช้จำแนกบทความทั้งหมดที่เป็นรายงานการวิจัยที่มีหัวข้อคล้ายคลึงกัน โดยยึดตามค่า Term frequency-inverse document frequency (TF-IDF) ของในแต่ละบทความ โดยการประเมินผลจากค่า F-Score สูงสุดคือการทำ TFIDF-LDA_30 โดยมี 15 คำสำคัญ และ 15 หัวข้อ

Wexler, A., Davoudi, A., Weissenbacher, D., Choi, R., Oconnor, K., Cummings, H., & Gonzalez, G. [14] ได้เสนอการวิเคราะห์ฟอรัมการตั้งครรภ์ออนไลน์ เพื่อให้เข้าใจวิธีการหาข้อมูลของหญิงตั้งครรภ์ในชุมชนออนไลน์ โดยใช้การสร้างแบบจำลองหัวข้อ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) โดยอัลกอริทึมนี้สามารถค้นพบ ถุงคำ (Bag of words) ที่มีความน่าจะเป็นสูงที่จะปรากฏร่วมกัน โดยผลลัพธ์ในการจัดหมวดหมู่ ได้หัวข้อที่ใหญ่ที่สุด ได้แก่ สุขภาพมารดา (45%) หัวข้อเกี่ยวกับทารก (29%) และ คน/ความสัมพันธ์ (10%) และในการจัดหมวดหมู่ตามไตรมาสการตั้งครรภ์ ผลลัพธ์ในไตรมาสแรกหญิงตั้งครรภ์มีความกังวลในเรื่องการแท้งบุตร ไตรมาสที่สามมีความกังวลในการคลอดบุตร และกิจวัตรการนอนของทารกในช่วงหลังคลอด

Twinandilla, S., Adhy, S., Surarso, B., & Kusumaningrum, R. [12] ได้เสนอการสรุปเอกสารหลายฉบับเพื่อให้ผู้อ่านสามารถเข้าใจเจตนาของเอกสารข่าวออนไลน์ได้อย่างง่ายดาย โดยการใช้เทคนิคการสร้างแบบจำลองหัวข้อ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ในการสรุปประโยคที่สำคัญในเอกสารหลายๆฉบับในภาพรวม โดยไม่ต้องจัดกลุ่มตามหัวข้อ และได้เสนอวิธีการสรุปแบบใหม่คือการรวม K-Means Clustering และ LDA - Significance Sentences เข้าด้วยกัน

Lim, David S. [8] ได้ทำการพัฒนาระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) สุขภาพจิต ชื่อว่า Woebot สำหรับคาดเดาผู้ใช้งานในเรื่องความรุนแรงของภาวะซึมเศร้า โดยป้อนข้อมูลไปยังตัวแยกประเภท (Classifier) 3 ประเภท ได้แก่ ข้อมูลการสำรวจพื้นฐาน หรือ เวกเตอร์ TFIDF จากข้อความทั้งหมด หรือเวกเตอร์การหาความน่าจะเป็นจากแบบจำลอง LDA โดยงานวิจัยนี้พบว่าหัวข้อที่สร้างโดย LDA มีประสิทธิภาพที่ดีสำหรับแสดงลักษณะอารมณ์จากข้อความ ซึ่งสามารถใช้เพื่อทำนายผลภาวะซึมเศร้า

Y. B. Touimi, A. Hadioui, N. E. Faddouli, and S. Bennani [11] บทความนี้กล่าวถึง ระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) ในเชิงความหมาย สำหรับการวิเคราะห์ แลกเปลี่ยน แบ่งปันความรู้และประสบการณ์ของผู้เรียน โดยงานวิจัยนี้ได้รับข้อความจากผู้ใช้งานด้วยระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) จำแนกคำที่มีนัยสำคัญด้วยแบบจำลอง LDA ทำการmapping ไปยังโดเมน ontology ของ MOOC การประยุกต์ใช้แบบจำลองความน่าจะเป็นของ LDA ทำให้สามารถดึงความรู้ที่เกี่ยวข้องตามที่ผู้เรียนร้องขอได้

จากงานวิจัยที่ได้กล่าวมานั้นเป็น การนำเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มาใช้ในการทำนายหัวข้อ ลดมิติของข้อมูลและการหาคำที่มีนัยสำคัญในเอกสารที่มีจำนวนมาก มีการนำไปประยุกต์ใช้กับการจัดกลุ่มเพื่อเพิ่มประสิทธิภาพในการจำแนกหมวดหมู่หัวข้อ ในด้านหาข้อเท็จจริงของข่าวสาร การจำแนกเอกสาร และการจัดหมวดหมู่ของเอกสาร รวมไปถึงการนำการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มาประยุกต์ใช้กับ ระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) ในการวิเคราะห์หาคำที่มีนัยสำคัญในข้อความ

วิธีดำเนินการ

ขั้นตอนที่ 1 : แนะนำชุดข้อมูลที่ใช้ในการศึกษานี้

ผู้วิจัยนำเข้าข้อมูลเกี่ยวกับคำปรึกษาปัญหาการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิดภาษาไทยสำหรับการทำวิจัยในครั้งนี้ โดยใช้ชุดข้อมูลในช่วงวันที่ 23 มกราคม 2017 ถึง 15 ตุลาคม 2020 จำนวนข้อมูลทั้งหมดมี 9,987 ข้อความ ในการเก็บรวบรวมข้อมูลได้ใช้ Selenium ในการทำ Web Scraping ข้อมูล ดึงข้อมูลจากเว็บไซต์ Honestdocs [4] ในหมวดหมู่คำปรึกษาปัญหาของการตั้งครรภ์ บันทึกข้อมูลในรูปแบบของ ไฟล์นามสกุล csv

ขั้นตอนที่ 2 : การจัดเตรียมข้อมูล (Data preparation)

ในการจัดเตรียมข้อมูลได้ใช้ ภาษาไพทอน (Python) ในการทำความสะอาดข้อมูล (Text cleaning) ตัดประโยคที่ไม่เกี่ยวข้องกับข้อความออกจากข้อความที่เกิดจากการดึงข้อมูลและตัดตัวอักษรพิเศษ (Special characters) กระบวนการตัดคำ (Word tokenize) นำข้อความทั้งหมดเข้าสู่กระบวนการตัดคำ โดยใช้ชุดคำสั่ง newmm และตัด Stop words ในการกำจัดคำที่ไม่จำเป็นหรือไม่เกี่ยวข้องออกจากข้อความ โดยใช้ชุดคำสั่งของ PythaiNLP โดยหลังจากการ เตรียมข้อมูล ข้อมูลคงเหลือ 8,248 ข้อความ

ขั้นตอนที่ 3 : การสกัดใจความสำคัญของข้อความ (Vectorizer)

กระบวนการสกัดใจความสำคัญจากข้อความ โดยกระบวนการนี้ใช้ชุดโปรแกรม Sklearn ชื่อว่า Tfidfvectorizer ในการสกัดคุณลักษณะ (Feature extraction) เพื่อหาความถี่ของคำที่มีลักษณะเฉพาะในแต่ละเอกสาร ช่วยในการเลือกคำที่มีนัยสำคัญในเอกสาร ผลลัพธ์ที่ได้จากการสกัดคุณลักษณะจากข้อความทั้งหมด 8,248 ข้อความ โดยมีการกำหนดตัวแปร max_df = 0.5 คือละเว้นคำที่ปรากฏในเอกสารมากกว่า 50% และ min_df = 10 คือละเว้นคำที่ปรากฏในเอกสารน้อยกว่า 10 ฉบับ โดยมีการนับจำนวนคำที่มีนัยสำคัญในแต่ละเอกสารทั้งหมดและนับจำนวนคำที่มีลักษณะเฉพาะ (Unique Words)

เมื่อเข้าสู่กระบวนการ Tfidfvectorizer จะได้ถุงคำ (Bag of word) ที่มีการนำความถี่ของคำที่มีลักษณะเฉพาะ (Unique Words) ที่ปรากฏในแต่ละเอกสาร โดยจำนวนคุณลักษณะ (Feature) จะขึ้นอยู่กับจำนวนของคำที่มีลักษณะเฉพาะ (Unique Words) โดยตัวอย่างของ ถุงคำ (Bag of word) ประกอบด้วย รหัสของเอกสาร รหัสคำที่ปรากฏในเอกสาร และความถี่ของการปรากฏคำในเอกสารนั้นๆ จากข้อความคำปรึกษาทั้งหมด มีจำนวนคำทั้งหมด 37,572 คำ และ จำนวนคำที่มีลักษณะเฉพาะ (Unique Words) จำนวน 414 คำ

ภาพที่ 1 การจับกลุ่มคำ (Word Cloud)

จากการจับกลุ่มคำ (Word Cloud) จากข้อมูลที่ทำการจัดเตรียมข้อมูลเรียบร้อยแล้ว โดยใช้ wordcloud matplotlib มีการนับจำนวนคำ ได้ผลลัพธ์ 10 อันดับของคำที่มีการกล่าวถึงมากที่สุด ได้แก่ ประจำเดือน, ตั้งครรภ์, มีเพศสัมพันธ์, มีโอกาส, ฉุกเฉิน, คุณหมอ, ถุงยาง,ป้องกัน, ยาคุมกำเนิด และครั้งแรก เป็นต้น โดยเมื่อพิจารณาจากการจับกลุ่มคำ จากภาพที่ 1 ผู้วิจัยได้ทำการพิจารณาและแบ่งกลุ่มของหัวข้อเป็น 4 หัวข้อดังนี้

(1) อาการผิดปกติในขณะตั้งครรภ์

(2) ความผิดปกติของประจำเดือน

(3) ความกังวลเมื่อมีเพศสัมพันธ์

(4) การคุมกำเนิด

ขั้นตอนที่ 4 : การสร้างแบบจำลองการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation)

เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ใช้เทคนิคนี้เพื่อต้องการจัดกลุ่มข้อความที่มีลักษณะคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน โดยการลดมิติของข้อมูล (Dimensionality-reduction) แนวคิดของเทคนิคนี้คือการกำหนดค่า m (จำนวนหัวข้อ) ไปยังในแต่ละ N (เอกสาร) สามารถหาเวกเตอร์ความยาวของค่า m ซึ่งแทนค่าด้วยการ กระจายความน่าจะเป็นของหัวข้อในแต่ละเอกสาร เพิ่มเวกเตอร์ไปยังเอกสารทั้งหมดจะได้ เมทริกซ์คุณลักษณะ N-by-m และมีการ Label หัวข้อในแต่ละเอกสาร

ขั้นตอนที่ 5 : การประเมินประสิทธิภาพของแบบจำลองการสร้างการจำแนกกลุ่มข้อความ (Clustering)

แนวคิดในการประเมินประสิทธิภาพของงานวิจัยนี้ เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) เป็นการสกัด feature ออกมาจาก bag of words (TF-IDF) โดยแต่ละ feature คือ คำที่กระจายตัวอยู่ในบางหัวข้อที่เป็นนามธรรม ซึ่งเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มีการวัดผลแบบ Intrinsic evaluation คือการใช้ metric ที่อยู่ในแบบจำลองเอง โดย เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มีการวัดผลโดยใช้ Perplexity หรือ ค่าความสับสนเพื่อดูการกระจายตัวของความน่าจะเป็นว่ามันดีมากน้อยเพียงใด แต่ Intrinsic Evaluation จะมีข้อเสียที่ว่า ในความจริง

แบบจำลองที่ได้นั้น มีประสิทธิภาพที่ดีกับงานจริงหรือไม่ เพราะว่าแบบจำลองถูกตั้งสมมุติฐานว่า ข้อมูลจริงหรือข้อมูลทดสอบ จะเหมือนกับชุดข้อมูลฝึกฝน ดังนั้นเราจึงใช้ K-mean ที่เป็นการวัดผลแบบ Extrinsic evaluation ด้วยการดูว่า vector ที่ได้จาก LDA มันให้ผลลัพธ์ได้ดีแค่ไหนใน downstream task

การประเมินประสิทธิภาพในการจำแนกกลุ่มข้อความ (Clustering) ใช้การประเมินประสิทธิภาพโดยการ วัดค่าในกลุ่มเดียวกันของสมาชิกในกลุ่มนั้นๆ และค่าความเหมือนระหว่างกลุ่มมากน้อยเพียงใด การคำนวณค่าได้ใช้ Silhouette Coefficient โดยชุดโปรแกรม Sklearn ที่ชื่อว่า silhouette_score ดังสมการ

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

โดย a คือ ระยะทางเฉลี่ยระหว่างจุดในกลุ่มเดียวกัน

b คือ ระยะทางเฉลี่ยระหว่างจุดระหว่างกลุ่ม

i คือ จุดของข้อมูล

ค่า Silhouette score S(i) มีค่าอยู่ในช่วง [-1, 1] ถ้ามีค่าใกล้ 1 มากคือดี

ผลการวิจัยและอภิปรายผลการวิจัย

งานวิจัยนี้ได้ทำการวิเคราะห์หาคำที่มีนัยสำคัญและการจำแนกกลุ่มหัวข้อในคำปรึกษาปัญหาการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิดในข้อความภาษาไทย โดยการใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) พัฒนาด้วยภาษาไพทอน (Python) และเครื่องมือ Google Colab มีการรวบรวมคำปรึกษาทั้งหมด 8,248 ข้อความ โดยมีการประเมินประสิทธิภาพของการจำแนกกลุ่มข้อมูล โดยการใช้ K-means และวัดผลด้วย Silhouette Coefficient และผลลัพธ์ของการจำแนกกลุ่มข้อความด้วยเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation)

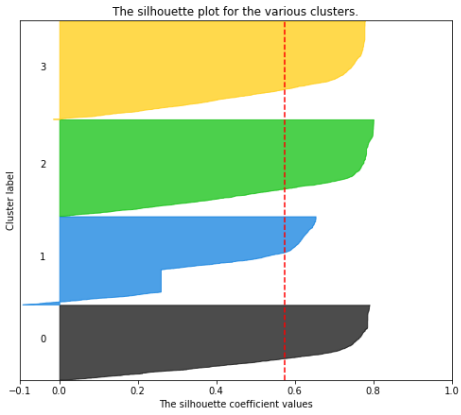
1. ผลลัพธ์การทดลองการจัดกลุ่มด้วยอัลกิริทึม K-mean และประเมินประสิทธิภาพการจำแนกกลุ่มด้วย Silhouette Coefficient

การจำแนกกลุ่มหัวข้อ โดยใช้ K-means ในการจำแนกกลุ่ม (Cluster) และประเมินประสิทธิภาพโดยใช้ Silhouette score ในการหาค่าจำนวนการจำแนกกลุ่มหัวข้อที่เหมาะสม ได้ผลลัพธ์ดังนี้

ตาราง 1 การประเมินประสิทธิภาพการจำแนกกลุ่มหัวข้อ

จำนวนคลัสเตอร์	ผลการประเมินประสิทธิภาพ
4	0.5397
5	0.4984
6	0.5174
7	0.410
8	0.4838
9	0.4424
10	0.4893

จากการประเมินประสิทธิภาพการจำแนกกลุ่มหัวข้อใน ตารางที่ 4 ผลลัพธ์ของการจำแนกกลุ่มหัวข้อ โดยใช้ K-means โดยการเปรียบเทียบการจำแนกคลัสเตอร์ 2 ถึง 10 คลัสเตอร์ พบว่า การจำแนกกลุ่มข้อมูล เป็น 4 หัวข้อนั้น ค่า Silhouette score มีค่า 0.5397 มีค่าเฉลี่ยในการแบ่งคลัสเตอร์ที่สูงที่สุด เมื่อเปรียบเทียบกับคลัสเตอร์อื่นๆ



ภาพที่ 2 กราฟ Silhouette score ในแต่ละหัวข้อ

จากภาพที่ 2 แสดงกราฟ Silhouette วัดค่าของความใกล้เคียงกันของจุดในคลัสเตอร์เดียวกันและวัดกับจุดในคลัสเตอร์อื่นๆ โดยจากกราฟ Silhouette วัดค่าของความใกล้เคียงกันของจุดในคลัสเตอร์เดียวกันและวัดกับจุดในคลัสเตอร์อื่นๆ โดยจากภาพในแกน x คือ ค่า Silhouette ในแกน y คือ จำนวนคลัสเตอร์ ซึ่งมีกลุ่มข้อมูล 4 คลัสเตอร์ โดยขนาดความกว้างในแนวแกน y บงชี้ถึงจำนวนของข้อความที่ประกอบอยู่ในหัวข้อนั้นๆ และเส้นประสีแดง คือ ค่าเฉลี่ยของค่า Silhouette score โดยเมื่อเปรียบเทียบจะเห็นว่า หัวข้อที่ 0, 3 และ 4 ประกอบด้วยข้อความใกล้เคียงกัน และมีค่า Silhouette score ที่ใกล้เคียงกัน และไม่มีคำที่ปะปนไปยังกลุ่มหัวข้ออื่นๆ หรือมีการปะปนเพียงเล็กน้อย และในหัวข้อที่ 1 มีการปะปนของกลุ่มข้อความอยู่เพียงเล็กน้อย

โดยดูจากค่าที่ติดลบของ Silhouette score และในหัวข้อที่ 1 ยังมีกลุ่มคำที่ต่ำกว่าค่าเฉลี่ย ซึ่งกราฟชี้ให้เห็นว่าการจับกลุ่มของข้อมูลนั้นในแต่ละคลัสเตอร์มีการจับกลุ่มที่ดีแต่ก็มีการปะปนของข้อความเล็กน้อย

2. ผลลัพธ์ของการจำแนกกลุ่มข้อความด้วยเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation)

ตารางที่ 2 ผลการวิเคราะห์หา 10 อันดับ คำที่ปรากฏในหัวข้อและหัวข้อของแบบจำลอง

ลำดับหัวข้อ	10 อันดับ คำที่ปรากฏในหัวข้อ	จำนวนข้อความที่ถูกทำนาย (ข้อความ)
1	('ประจำเดือน', 1100.144), ('มีเพศสัมพันธ์', 600.697), ('ฉุกเฉิน', 389.811), ('มีโอกาส', 386.944), ('ป้องกัน', 278.269), ('ถุงยาง', 256.189), ('ครั้งแรก', 138.034), ('แล้วก็', 120.891), ('แบบนี้', 120.824), ('หลังจากนั้น', 108.507)	3156
2	('ตั้งครรภ์', 512.758), ('คุมกำเนิด', 102.379), ('ค่าใช้จ่าย', 86.645), ('ที่ผ่านมา', 81.052), ('รอบเดือน', 74.40), ('อันตราย', 72.434), ('สีน้ำตาล', 71.968), ('คุณหมอ', 63.779), ('ประจำเดือน', 61.396), ('ฝากครรภ์', 60.656)	1681
3	('ยาคุมกำเนิด', 213.260), ('คุณหมอ', 180.215), ('ประจำเดือน', 140.011), ('ตั้งครรภ์', 106.0836), ('แบบนี้', 105.248), ('ปวดท้อง', 82.235), ('โรงพยาบาล', 75.392), ('กลับมา', 56.274), ('ฮอร์โมน', 55.680), ('รับประทาน', 50.166)	2079
4	('ทำหมัน', 124.553), ('ช่องคลอด', 121.687), ('เข้าไป', 114.509), ('สอดใส่', 110.712), ('อวัยวะเพศ', 101.148), ('พันธุ์', 80.968), ('มีโอกาส', 79.538), ('หล่อลื่น', 78.555, ('น้ำอสุจิ', 69.058), ('ท้องน้อย', 67.856)	1332

ผลการทำนายของแบบจำลองได้วิเคราะห์คำที่มีนัยสำคัญและหัวข้อ ในตารางที่ 2 พบว่า แบบจำลองหัวข้อได้จำแนกหัวข้อให้กับข้อความทั้งหมด 8,248 ข้อความ โดยในคอลัมน์ที่ 2 เป็น 10 อันดับของคำที่ปรากฏในหัวข้อในหัวนั้นๆ ซึ่งแต่ละหัวข้อมีการวิเคราะห์ชุดของคำที่ประกอบในหัวข้อและคำหลักแต่ละคำมีการถ่วงน้ำหนัก (weightage) กำกับให้กับแต่ละหัวข้อ ซึ่งเมื่อวิเคราะห์ผลลัพธ์ที่ได้ พบว่า

การทำนายว่าเป็นหัวข้อที่ 1 จำนวน 3156 ข้อความ โดยมีตัวอย่างของข้อความคำปรึกษาในหัวข้อที่ 1 มีตัวอย่าง ดังนี้

- กินยาคุมประจำเดือนไม่มาตรวจรอบภายในอาทิตย์กว่าๆ แล้วไม่ท้องสามารถเริ่มทานยาคุมแผงต่อไปได้มั้ยคะหรือต้องรอประจำเดือนมาก่อน
- หมอค่ะคือหนูมีอะไรกับแฟนเมื่อวานไม่ได้ป้องกันและไม่แน่ใจว่าแฟนดึงออกทันมั้ยถ้าจะกินยาคุมแบบเม็ดวันนี้จะกินทันไหมคะ
- หมอคะหนูสงสัยค่ะคือหนูกินยาคุมแบบเม็ดเข้าแผงที่แล้วแต่มีเพศสัมพันธ์กับแฟนวันเดียวแล้วหลั่งใน อยากทราบว่าโอกาสในการตั้งครรภ์มีมากน้อยแค่ไหนคะ

การทำนายว่าเป็นหัวข้อที่ 2 จำนวน 1681 ข้อความ โดยมีตัวอย่างของข้อความคำปรึกษาในหัวข้อที่ 2 มีตัวอย่าง ดังนี้

- เล่นฮูลาฮูปในตอนที่ตั้งครรภ์อยู่ได้ไหมคะ จะทำให้แท้งไหมคะ
- คลอดลูกได้เดือนแล้วค่ะ ลืมทานยาแบบให้นมบุตรเดลิดอนเวลาทานตอนกลางคืนคะตอนเช้ามีเพศสัมพันธ์จะท้องไหมคะ ถ้าลืมเราต้องทำอย่างไรกังวลมากเลยค่ะ
- ตั้งครรภ์อยู่ตรวจเลือดเจอพาหะกามโรคจะมีผลต่อเด็กในครรภ์ไหมคะ

การทำนายว่าเป็นหัวข้อที่ 3 จำนวน 2079 ข้อความ โดยมีตัวอย่างของข้อความคำปรึกษาในหัวข้อที่ 3 มีตัวอย่าง ดังนี้

- กินยาคุมมาเดือน แล้วกินเผื่อพลาดหยุดได้ไหมคะจะท้องไหม
- กินยาคุมแอนนาอยู่ตั้งแต่วันที่กพแต่ตอนนี้เมนส์มาแล้วแต่ยังกินไม่หมดแผงแรกสามารถเริ่มกินแผงใหม่ได้เลยไหม
- หนูฝั่งเข็มยาคุมกำเนิดมาปีค่ะ ประจำเดือนไม่เคยมาเลยแต่ช่วงอาทิตย์ก่อนประจำเดือนมาค่ะแต่มาไม่เยอะเป็นเลือดสีดำน้ำตาลเป็นประมานวันก็หมดค่ะแบบนี้มีโอกาสท้องไหมคะ

การทำนายว่าเป็นหัวข้อที่ 4 จำนวน 1332 ข้อความ โดยมีตัวอย่างของข้อความคำปรึกษาในหัวข้อที่ 4 มีตัวอย่าง ดังนี้

- กินยาคุมประจำเดือนไม่มาตรวจรอบภายในอาทิตย์กว่าๆแล้วไม่ท้องสามารถเริ่มทานยาคุมแผงต่อไปได้มั้ยคะหรือต้องรอประจำเดือนมาก่อน
- หมอค่ะคือหนูมีอะไรกับแฟนเมื่อวานไม่ได้ป้องกันและไม่แน่ใจว่าแฟนดึงออกทันมั้ยถ้าจะกินยาคุมแบบเม็ดวันนี้จะกินทันมั้ยค่ะ
- หมอคะหนูสงสัยคะคือหนูกินยาคุมแบบเม็ดเข้าแผงที่แล้วแต่มีเพศสัมพันธ์กับแฟนวันเดียวมีรอบแล้วหลั่งในทั้งรอบอยากทราบว่าโอกาสในการตั้งครรภ์มีมากน้อยแค่ไหนคะ

แบบจำลองการวิเคราะห์คำที่มีนัยสำคัญและหัวข้อจากข้อความ มีจำนวนข้อความทั้งสิ้น ทั้งหมด 8,248 ข้อความ ซึ่งผู้วิจัยได้เปรียบเทียบคำสำคัญในแต่ละหัวข้อกับหัวข้อที่ผู้วิจัยตั้งสมมุติฐานไว้เบื้องต้น โดยสมมุติฐานที่ตั้งไว้เบื้องต้น ได้แ ก่ (1) อาการผิดปกติในขณะตั้งครรภ์ (2) ความผิดปกติของประจำเดือน (3) ความกังวลเมื่อมีเพศสัมพันธ์ (4) การคุมกำเนิด

ตาราง 3 ตารางผลลการเปรียบเทียบสมมุติฐานกับผลการทำนายของแบบจำลอง

สมมุติฐาน	1	2	3	4
อาการผิดปกติในขณะตั้งครรภ์		/	/	
ความผิดปกติของประจำเดือน		/	/	
ความกังวลเมื่อมีเพศสัมพันธ์	/			/
การคุมกำเนิด	/		/	/

โดยจากการตั้งสมมุติฐาน และวิเคราะห์คำที่มีนัยสำคัญในแต่ละหัวข้อ พบว่า

หัวข้อที่ 1 มีความเกี่ยวข้องกับความกังวลเมื่อมีเพศสัมพันธ์และการคุมกำเนิด โดยมีคำที่เกี่ยวข้องได้แก่ ประจำเดือน มีเพศสสัมพันธ์ ป้องกัน ถุงยาง และเครียด

หัวข้อที่ 2 มีความเกี่ยวข้องกับ อาการผิดปกติในขณะตั้งครรภ์และความผิดปกติของประจำเดือน โดยมีคำที่เกี่ยวข้องได้แก่ มีคำว่าตั้งครรภ์ คุมกำเนิด รอบเดือน อันตราย และสีน้ำตาล

หัวข้อที่ 3 มีความเกี่ยวข้องกับ อาการผิดปกติในการตั้งครรภ์ ความผิดปกติของประจำเดือนและการคุมกำเนิด โดยมีคำที่เกี่ยวข้อง ได้แก่ ยาคุมกำเนิด ประจำเดือน ตั้งครรภ์ และรัปประทาน

หัวข้อที่ 4 มีความเกี่ยวข้องกับ ความกังวลหลังมีเพศสัมพันธ์,ความผิดปกติของประจำเดือน มีคำที่เกี่ยวข้อง การทำหมัน สอดใส่ ประจำเดือน และผิดปกติ

เมื่อพิจารณาพบว่าการตั้งสมมุติฐานเบื้องต้นและผลการทำนาย พบว่า ผู้ใช้งานมีการเข้าปรึกษาปัญหาเกี่ยวกับ อาการผิดปกติในการตั้งครรภ์ ความผิดปกติของประจำเดือนและการคุมกำเนิด ซึ่งผลการทำนายว่าเป็นหัวข้อที่ 3 จำนวน 3020 ข้อความมากที่สุด นั่นหมายความว่าผู้ใช้งานส่วนมากที่เข้ามาปรึกษากับแพทย์นั้นมีความกังวลเรื่อง การทานยาคุม การคุมกำเนิด ความผิดปกติของประจำเดือน อาจจะเป็นจากการทานยาคุม หรือความกังวลหลังการมีเพศสัมพันธ์นั่นเอง และการพิจารณาพบว่าการตั้งสมมุติฐานเบื้องต้นไม่ได้กล่าวถึง การทำหมัน สิทธิบัตรทองหรือค่าใช้จ่าย ซึ่งการทำ วิธีการวิเคราะห์หาคำที่มีนัยสำคัญและหัวข้อในข้อความ ด้วยเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ทำให้เห็นคำที่แฝงอยู่ในข้อความ มากกว่าผู้วิจัยได้ตั้งสมมุติฐานไว้

สรุปผลการวิจัย

ปัจจุบันชุมชนออนไลน์เป็นพื้นที่หนึ่ง ที่ทำให้คนจำนวนมากสามารถเข้าปรึกษาแพทย์ผู้เชี่ยวชาญได้สะดวก แต่ผู้ใช้งานต้องรอการตอบกลับจากแพทย์ เพราะแพทย์ไม่สามารถให้คำปรึกษาได้ทันทีเมื่อผู้ใช้งานขอคำปรึกษา (Availability) และหากมีปริมาณคำปรึกษามาก แพทย์ไม่สามารถ ให้คำปรึกษาได้ทัน (Scalability) ซึ่งจากข้อจำกัดเหล่านี้ ระบบโต้ตอบอัตโนมัติสามารถแก้ไขได้ แต่ว่าระบบโต้ตอบอัตโนมัติยังคงมีความท้าทายในด้านภาษาธรรมชาติ ที่ต้องทำให้คอมพิวเตอรสามารถเข้าใจภาษามนุษย์

งานวิจัยนี้เป็นการศึกษาวิธีการวิเคราะห์หาคำที่มีนัยสำคัญและหัวข้อในข้อความและทำการจัดกลุ่มข้อความที่มีลักษณะคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน ด้วยเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ชี้ให้เห็นว่าผู้ใช้งานอินเตอร์เน็ต

จำนวนมากที่เข้าปรึกษาปัญหากับแพทย์ผ่านชุมชนออนไลน์นั้น ส่วนมากมีความกังวลในเรื่องเรื่อง เกี่ยวกับ อาการผิดปกติในการตั้งครรภ์ ความผิดปกติของประจำเดือนและการคุมกำเนิด โดยงานวิจัยนี้ได้ทำการแบ่งกลุ่มข้อมูลโดยใช้ K-means ในการวัดประสิทธิภาพในการจัดกลุ่มของ เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) โดยค่า Silhouette Coefficient มีการจัดกลุ่มข้อมูลที่ 4 กลุ่ม มีค่าอยู่ที่ 0.5397 ซึ่ง มีค่าที่สูงและข้อมูลมีการจับกลุ่มได้ดีเมื่อเทียบกับการจัดกลุ่มด้วยจำนวนอื่นๆ

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] A. A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B. M. Bewick, and M. Househ, "Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review," (in English), *J Med Internet Res,* Review vol. 23, no. 1, p. e17828, 2021, doi: 10.2196/17828.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[3] A. Chaves and M. Gerosa, "How Should My Chatbot Interact? A Survey on Social Characteristics in HumaŽ Chatbot Interaction Design," *International Journal of Human Computer Interaction,* vol. 37, pp. 729 - 758, 2021.

[4] H. E. Department, vol. 2020, no. 21 November. [Online]. Available: https://hd.co.th/

[5] S. H., "Text mining Twitter social media for Covid-19: Comparing latent semantic analysis and latent Dirichlet allocation," 2020. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-32567.

[6] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications,* vol. 78, no. 11, pp. 15169-15211, 2019/06/01 2019, doi: 10.1007/s11042-018-6894-4.

[7] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences,* vol. 9, no. 1, p. 30, 2019/08/26 2019, doi: 10.1186/s13673-019-0192-7.

[8] D. S. Lim, "Predicting outcomes in online chatbot-mediated therapy," ed, 2017.

[9] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, "A Survey on Evaluation Methods for Chatbots," presented at the Proceedings of the 2019 7th International Conference on Information and Education Technology, Aizu-Wakamatsu, Japan, 2019. [Online]. Available: https://doi.org/10.1145/3323771.3323824.

[10] H. Raval, "Limitations of Existing Chatbot with Analytical Survey to Enhance the Functionality Using Emerging Technology," *International Journal of Research and Analytical Reviews (IJRAR),* vol. 7, no. 2, April 26, 2020 2020.

[11] Y. B. Touimi, A. Hadioui, N. E. Faddouli, and S. Bennani, "Intelligent Chatbot-LDA Recommender System," *International Journal of Emerging Technologies in Learning (iJET),* vol. 15, no. 20, 2020. [Online]. Available: https://www.scilit.net/article/d4fac79b4922dceeb80df6da6deaa715.

[12] S. Twinandilla, S. Adhy, B. Surarso, and R. Kusumaningrum, "Multi-Document Summarization Using K-Means and Latent Dirichlet Allocation (LDA) – Significance Sentences," *Procedia Computer Science,* vol. 135, pp. 663-670, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.procs.2018.08.220.

[13] G. K. Vamsi, A. Rasool, and G. Hajela, "Chatbot: A Deep Neural Network Based Human to Machine Conversation Model," in 2020 11*th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-3 July 2020 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225395.

[14] A. Wexler *et al.*, "Pregnancy and health in the age of the Internet: A content analysis of online "birth club" forums," *PLOS ONE,* vol. 15, p. e0230947, 04/14 2020, doi: 10.1371/journal.pone.0230947.