กรอบการวิเคราะห์รูปแบบการลงคะแนนเสียงเลือกพรรคการเมืองในกลุ่มผู้ใช้ Twitter ตามผลการทำนายความรู้สึกโดยใช้สถาปัตยกรรม BERT

กานติมา เตชะผลประสิทธิ์*[1]

คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

โอม ศรนิล[2]

คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

** Corresponding Author: kantima.tec@stu.nida.ac.th*

[1] นักศึกษาระดับปริญญาโท สาขาวิชาการวิเคราะห์ข้อมูลและวิทยาการข้อมูล คณะสถิติประยุกต์

[2] รองศาสตราจารย์ คณะสถิติประยุกต์

บทคัดย่อ

ปัจจุบันผู้คนส่วนใหญ่เลือกที่จะแสดงความคิดเห็นทางการเมืองผ่านทางแพลตฟอร์มทวิตเตอร์ ไม่ว่าจะเป็นการโพสต์ข้อความ (Tweet), การกดถูกใจ (Like) และการแชร์ข้อความ (Retweet) ด้วยเหตุนี้จึงเป็นที่มาของการศึกษานี้ว่าความคิดเห็นบนช่องทางออนไลน์บนแพลตฟอร์มทวิตเตอร์สอดคล้องกับผลการเลือกตั้งทั่วไปแบบบัญชีรายชื่อ ปี พ.ศ.2566 จริงหรือไม่ และเพื่อศึกษาว่าผู้ใช้ทวิตเตอร์ชื่นชอบหรือโจมตีพรรคการเมืองใด รวมถึงศึกษาหาความสัมพันธ์ระหว่างอันดับผลการเลือกตั้งรายพรรคการเมืองกับปัจจัยข้อมูลต่างๆบนTwitter เช่น จำนวนTweet, จำนวนการกด Like, จำนวนการ Retweet โดยผู้วิจัยได้ทำการเปรียบเทียบ Model WangchanBERTa กับ Logistic Regression เพื่อหา Modelที่มีประสิทธิภาพสูงสุดในการจำแนกความรู้สึกของข้อความ เพื่อนำไปใช้ในกรอบการวิเคราะห์รูปแบบการลงคะแนนเสียงเลือกพรรคการเมืองในกลุ่มผู้ใช้ Twitter โดยผลการวิจัยพบว่า Model WangchanBERTa ให้ประสิทธิภาพสูงสุดในการจำแนกความรู้สึกของข้อความ และจากกรอบการวิเคราะห์พบว่า ผู้ใช้ทวิตเตอร์ส่วนใหญ่ชื่นชอบพรรคเพื่อไทย และโจมตีพรรครวมไทยสร้างชาติ และปัจจัยที่สำคัญที่มีผลต่ออันดับผลการเลือกตั้งคือ จำนวนการ Retweet ในความเห็นเชิงบวกและจำนวนการ Retweet ในความเห็นที่เป็นกลาง

คำสำคัญ : Twitter Analysis /Sentiment Analysis / WangchanBERTa

Framework for Analyzing Party Voting Patterns among Twitter Users Based On Sentiment Prediction Results Using BERT Architecture

Kantima Techaphonprasit*[1]

Faculty of Applied Statistics, National Institute of Development Administration

Ohm Sornil[2]

Faculty of Applied Statistics, National Institute of Development Administration

** Corresponding Author: kantima.tec@stu.nida.ac.th*

[1] Master's Student, Department of Data Analytics and Data Science, Faculty of Applied Statistic

[2] Associate Professor, Department of Computer Science, Faculty of Applied Statistic

Abstract

This study delves into the relationship between political sentiments expressed on Twitter and their alignment with the party-list system's overall results in the 2566 B.E. general elections. It seeks to uncover Twitter users' preferences for specific political parties and their inclination to criticize others. Additionally, the investigation explores the correlation between election rankings of political parties and essential Twitter data factors, such as Tweet frequency, Likes, and Retweets.

To achieve these goals, we compared the effectiveness of two sentiment classification models, WangchanBERTa and Logistic Regression, for analyzing textual data's sentiment. The chosen model, WangchanBERTa, proved to be more efficient, thus being integrated into a comprehensive framework for analyzing party voting patterns among Twitter users.

The study findings reveal Twitter's significant role as a platform for political expression, with a majority of users actively sharing their views through textual content, Likes, and Retweets. The WangchanBERTa model showcased remarkable performance in sentiment classification, making it a valuable tool for analyzing political discussions on Twitter.

The analysis framework sheds light on Twitter users' strong preference for the Pheu Thai Party and their notable criticism towards the Ruam Thai Sang Chad Party. Moreover, the number of Retweets in positive and neutral sentiments emerged as crucial factors influencing election rankings.

By linking online sentiments to real-world election outcomes, this study contributes valuable insights to political analysis. Furthermore, it underscores the importance of utilizing Twitter data as a predictive tool to better understand voting patterns and political preferences among its user base.

Keywords: Twitter Analysis /Sentiment Analysis / WangchanBERTa

1. บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

การเกิดขึ้นของแพลตฟอร์มโซเชียลมีเดียได้ปฏิวัติภูมิทัศน์ของวาทกรรมทางการเมือง โดยเปิดโอกาสให้บุคคลได้แบ่งปันความคิดเห็น มีส่วนร่วมในการอภิปราย และแสดงความคิดเห็นเกี่ยวกับปัญหาต่างๆ ของสังคม รวมถึงการเมือง โดยเฉพาะ Twitter ได้กลายเป็นแพลตฟอร์มที่โดดเด่นสำหรับผู้ใช้ในประเทศไทยในการแสดงความคิดเห็นและมีส่วนร่วมในการสนทนาทางการเมือง ด้วยลักษณะเรียลไทม์และความง่ายในการใช้งาน Twitter ได้ดึงดูดชุมชนผู้ใช้ที่หลากหลาย ซึ่งรวมถึงนักการเมือง นักเคลื่อนไหว ผู้สื่อข่าว และพลเมืองทั่วไป ซึ่งทั้งหมดนี้มีบทบาทสำคัญในการกำหนดการเล่าเรื่องทางการเมืองในประเทศ

ในช่วงไม่กี่ปีที่ผ่านมา การใช้ Twitter เป็นสื่อกลางในการแบ่งปันความคิดเห็นทางการเมืองในประเทศไทยมีการเติบโตแบบทวีคูณ แพลตฟอร์มนี้จึงกลายเป็นแหล่งข้อมูลเรียลไทม์ที่สำคัญที่สะท้อนถึงความรู้สึกของสาธารณะและทางเลือกทางการเมือง นักวิจัยและนักวิเคราะห์หันมาใช้ข้อมูล Twitter กันมากขึ้นเพื่อหาข้อมูลเชิงลึกเกี่ยวกับรูปแบบการโหวต ความนิยมของพรรคการเมือง และประสิทธิภาพของนโยบายของรัฐบาล

การศึกษานี้จึงมุ่งที่จะเจาะลึกเข้าไปในขอบเขตของความคิดเห็นทางการเมืองที่แบ่งปันบน Twitter ในประเทศไทย ตั้งแต่วันที่ 1 มกราคม พ.ศ.2566 จนถึงวันที่ 10 เมษายน พ.ศ.2566 เป็นเวลารวม 100 วัน หรือเป็นช่วงเวลาที่พรรคการเมืองต่างๆกำลังหาเสียงก่อนวันเลือกตั้งในวันที่ 14 พฤษภาคม พ.ศ.2566 โดยรวบรวมข้อมูลจำนวนทั้งสิ้น 102,997 ข้อความ แต่แค่จำนวนข้อความ (No. of Tweet) เพียงอย่างเดียวอาจไม่ได้สะท้อนความคิดเห็นที่แท้จริง จึงได้มีการนำ Model เข้ามาช่วยในการจำแนกความคิดเห็นของแต่ละข้อความว่าเป็นไปในเชิงบวก เชิงลบ หรือกลางๆ เพื่อทำให้สามารถวิเคราะห์ได้ลึกและหลากหลายมิติมากขึ้น โดยจะต้องทำการคัดเลือก Model ที่มีประสิทธิภาพสูงสุดในการจำแนกความรู้สึกของข้อความภาษาไทย จากนั้นจึงนำผลลัพธ์ที่ได้ไปทำการวิเคราะห์ตามกรอบการวิเคราะห์ (Framework) ต่อไป

1.2 วัตถุประสงค์ของการวิจัย

1) เพื่อศึกษาความสัมพันธ์ระหว่างความคิดเห็นบนโลกออนไลน์บนแพลตฟอร์ม Twitter กับผลการเลือกตั้งทั่วไปแบบบัญชีรายชื่อปี พ.ศ.2566 ว่ามีความสอดคล้องกันหรือไม่ และมีความสัมพันธ์กันอย่างไร โดยใช้ Model BERT (Bidirectional Encoder Representations from Transformers) ในการจำแนกความรู้สึกของข้อความ Twitter

2) เพื่อศึกษาว่าผู้ใช้ Twitter ชื่นชอบหรือโจมตีพรรคการเมืองใด

3) เพื่อศึกษาหาความสัมพันธ์ระหว่างอันดับผลการเลือกตั้งรายพรรคการเมืองกับปัจจัยข้อมูลต่างๆบน Twitter เช่น จำนวน Tweet, จำนวนการกดLike, จำนวนการ Retweet และความรู้สึกของข้อความ

2. แนวคิดและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาแนวคิดและงานวิจัยที่เกี่ยวข้องกับการทำนายผลการเลือกตั้งโดยการวิเคราะห์ความรู้สึกของข้อความจากข้อมูลใน Twitter ค่อนข้างมีความหลากหลาย แต่เท่าที่พบส่วนใหญ่เป็นการวิจัยของต่างประเทศและจะใช้ข้อมูลตั้งแต่หลักหมื่นขึ้นไปจนถึงหลักแสนข้อความ และมีค่าความแม่นยำอยู่ที่ประมาณ 70% ขึ้นไป โดย Model ที่ใช้ในการจำแนกความรู้สึกมีหลากหลาย เช่น BERT, Support Vector Machine (SVM), Logistic Regression หรือ Naive Bayes เช่น งานวิจัยของ [1] เป็นการวิเคราะห์ความรู้สึก (Sentiment analysis) ข้อความบน Twitter ที่เกี่ยวกับการเลือกตั้งทั่วไปของประเทศเยอรมันในปี ค.ศ. 2021 โดยการเก็บข้อมูลจาก Twitter ทั้งสิ้น 58,000 ข้อความจากทั้งหมด 7 พรรคการเมือง แล้วนำเข้า transformer-based models (BERT) ซึ่งมีความแม่นยำ (Accuracy) อยู่ที่ 93% หรืองานวิจัยของ [2] ได้ทำการเก็บข้อมูลภาษาฮินดีจาก Twitter ทั้งสิ้น 42,235 ข้อความ เป็นเวลา 1 เดือนในช่วงการหาเสียงเลือกตั้งทั่วไปของประเทศอินเดีย โดยใช้เทคนิค Dictionary Based และ Model Naive Bayes และ SVM ในการจำแนกความรู้สึกของข้อความ ซึ่ง SVM ให้ค่าความแม่นยำสูงสุดที่ 78.4% โดยผู้เขียนได้วิเคราะห์เพิ่มเติมว่า ข้อมูล

Twitter หลักหมื่นอาจไม่เพียงพอที่จะ train model หรืองานวิจัยของ [3] ที่น่าสนใจเช่นกัน เป็นการวิเคราะห์ความรู้สึกของข้อความ Twitter ในช่วงการหาเสียงเลือกตั้งประธานาธิบดีของประเทศฟิลิปปินส์ในปี ค.ศ. 2022 โดยทำการเก็บข้อมูลทั้งสิ้นกว่า 114,000 ข้อความ และใช้ Multinomial Naïve Bayes model ในการจำแนกความรู้สึก ซึ่งมีความแม่นยำ (Accuracy) อยู่ที่ 84.8% และยังมีงานวิจัยของ [4] ได้ทำการวิเคราะห์ความรู้สึก โดยนำเข้าข้อความจาก Twitter ที่เกี่ยวกับการเลือกตั้งของประเทศจากาตาร์ โดยใช้เทคนิคของ Logistic Regression และวัดประสิทธิภาพด้วย Accuracy ซึ่งได้ค่าที่ 70% โดยกล่าวว่าจำนวนข้อมูลในการ train model มีความสำคัญมาก ซึ่งส่งผลกระทบโดยตรงต่อความแม่นยำของ Model ยิ่งมีข้อมูลมากก็จะทำให้ประสิทธิภาพ Model สูงขึ้น

สำหรับการวิเคราะห์ความรู้สึกของข้อความที่เป็นภาษาไทยได้มีการพัฒนามาอย่างต่อเนื่อง ซึ่ง Model ประมวลผลภาษาไทยที่มีขนาดใหญ่และก้าวหน้าที่สุดในขณะนี้คือ WangchanBERTa [5] โดย train language model สถาปัตยกรรม BERT ด้วยข้อมูลภาษาไทยขนาดใหญ่ที่มีความหลากหลายและถูกทำความสะอาดมากที่สุด โดยใช้กฎการจัดการข้อมูลที่สร้างขึ้นเพื่อภาษาไทยโดยเฉพาะ ซึ่ง WangchanBERTa ที่ใช้ในการจำแนกความรู้สึกใช้ชุดข้อมูลของ Wisesight-Sentiment [6] ซึ่งเป็นชุดข้อมูลที่ได้มาจากสื่อสังคมออนไลน์ (Social Media)

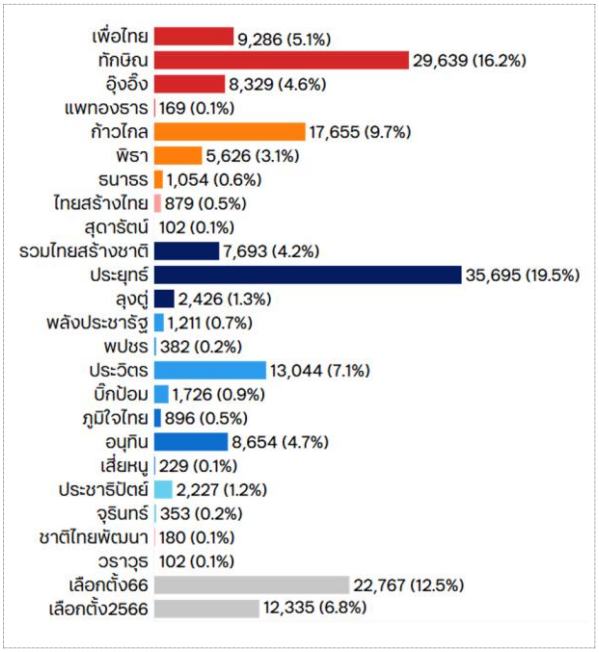
3. วิธีวิจัย

ในการวิเคราะห์รูปแบบการลงคะแนนเสียงเลือกพรรคการเมืองในกลุ่มผู้ใช้ Twitter ตามผลการทำนายความรู้สึกโดยใช้สถาปัตยกรรม BERT มีขั้นตอนการดำเนินงานตามรูปที่ 1 โดยเริ่มจากการรวบรวมข้อมูลจาก Twitter มาผ่านกระบวนการ Data Preprocessing และ Feature Extraction จากนั้นแบ่งข้อมูลเพื่อ Train/Test model แล้วจึงวัดประสิทธิภาพโมเดล เพื่อเลือกโมเดลที่มีประสิทธิภาพสูงสุดในการนำไปวิเคราะห์ต่อไปในกรอบการวิเคราะห์ (Framework)

รูปที่ 1 กระบวนการทำข้อมูล (Data Process)

3.1 การรวบรวมข้อมูล (Data Collection)

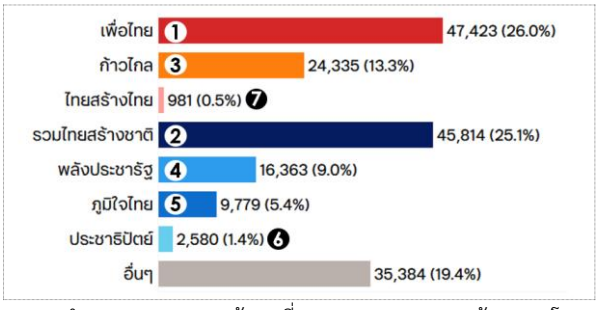
ใช้ SNscrape Library [7] ในการรวบรวมข้อมูลจาก Twitter โดยเก็บรวบรวมข้อมูลตั้งแต่วันที่ 1 มกราคม พ.ศ.2566 จนถึงวันที่ 10 เมษายน พ.ศ.2566 เป็นเวลารวม 100 วัน หรือเป็นช่วงเวลาที่พรรคการเมืองต่างๆกำลังหาเสียงก่อนวันเลือกตั้งในวันที่ 14 พฤษภาคม พ.ศ.2566 โดยรวบรวมข้อมูลจำนวนทั้งสิ้น 182,659 ข้อความ จากการค้นหาคำสำคัญ (Keywords) ที่เกี่ยวกับชื่อพรรคการเมืองหลักๆและชื่อนักการเมืองที่สำคัญจำนวน 25 คำตามรูปที่ 2 ดังนี้ 1) เพื่อไทย, 2) ทักษิณ, 3) อุ๊งอิ๊ง, 4) แพทองธาร, 5) ก้าวไกล, 6) พิธา, 7) ธนาธร, 8) ไทยสร้างไทย, 9) สุดารัตน์ ,10) รวมไทยสร้างชาติ, 11) ประยุทธ์, 12) ลุงตู่, 13) พลังประชารัฐ, 14) พปชร, 15) ประวิตร, 16) บิ๊กป้อม, 17) ภูมิใจไทย, 18) อนุทิน, 19) เสี่ยหนู, 20) ประชาธิปัตย์, 21) จุรินทร์, 22) ชาติไทยพัฒนา, 23) วราวุธ, 24) เลือกตั้ง66 และ25) เลือกตั้ง2566



รูปที่ 2 จำนวน Tweet และร้อยละของจำนวน Tweet จากข้อมูลทั้งหมดที่รวบรวมมา 182,659 ข้อความ โดยแบ่งตาม 25 คำค้นหา

ข้อมูลที่เก็บรวบรวมมี 6 ปัจจัย ดังนี้ 1) วันเวลาที่โพสต์ข้อความ 2) ชื่อผู้ใช้ (Username) 3) ข้อความที่โพสต์ 4) รหัสข้อความ (Tweet ID) 5) จำนวนการกดถูกใจ (Like) และ 6) จำนวนการแชร์ (Retweet)

เมื่อพิจารณาสัดส่วนข้อความแล้วพบว่าบางพรรคการเมืองถูกกล่าวถึงค่อนข้างน้อย ผู้วิจัยจึงเลือกที่จะวิเคราะห์เพียง 7 พรรคการเมืองหลัก คือ พรรคเพื่อไทย, พรรคก้าวไกล, พรรคไทยสร้างไทย, พรรครวมไทยสร้างชาติ, พรรคพลังประชารัฐ, พรรคภูมิใจไทย และพรรคประชาธิปัตย์ จากคำค้นหาทั้งหมด 25 คำ จึงถูกรวมเป็น 7 คำสำคัญตามชื่อพรรคการเมืองหลักตามรูปที่ 3



รูปที่ 3 จำนวน Tweet และร้อยละของจำนวน Tweet จากข้อมูลที่รวบรวมมา 182,659 ข้อความ โดยแบ่งตาม 7 พรรคการเมืองที่สนใจ

3.2 การเตรียมข้อมูล (Data Preprocessing)

เริ่มจากการทำความสะอาดข้อมูล โดยการลบข้อความซ้ำที่ถูกโพสต์โดยผู้ใช้ (Username) เดียวกันและรหัสข้อความ (Tweet ID) เดียวกัน จากนั้นเข้าสู่กระบวนการการทำความสะอาดข้อมูลประเภทข้อความ ดังนี้

3.2.1) ลบ Emoji

3.2.2) ลบสิ่งที่ไม่ใช่ตัวอักษร เช่น [], {}, %

3.2.3) ลบ URL

3.2.4) ลบข้อความด้านหลัง # (Hashtag)

3.2.5) ลบคำฟุ่มเฟือยภาษาไทย (Stop word) เช่น มี, การ, ความ เป็นต้น

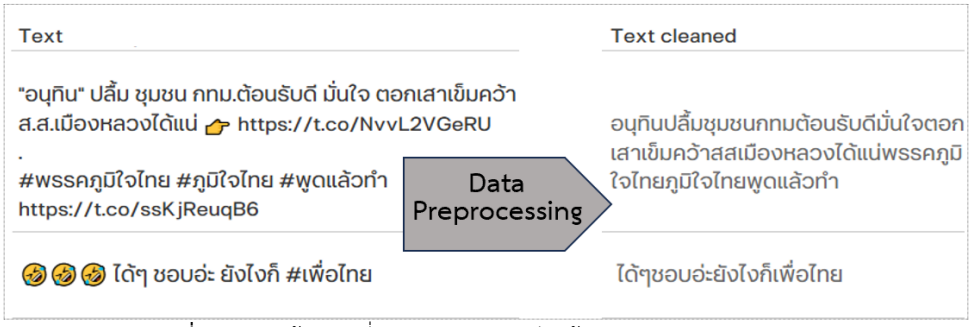
3.2.6) แก้ไขคำที่มีตัวอักษรซ้ำ เช่น “ดีมากกก” แก้ใช้เป็น “ดีมาก”

3.2.7) การตัดแบ่งคำเพื่อนำเข้า Model (Word Tokenized) โดยในงานวิจัยนี้ มีการใช้การตัดแบ่งคำ 3 แบบ ดังนี้

(1) newmm: การตัดแบ่งคำจาก dictionary ของคำในภาษาไทยด้วย maximal matching algorithm ด้วยไลบรารี่ PyThaiNLP [8]

(2) spm: การตัดแบ่งหน่วยคำย่อย (subword-level tokenization) ด้วยไลบรารี่ SentencePiece [9] โดยตัวตัดแบ่งหน่วยคำย่อยอาศัยข้อมูลทางสถิติของการปรากฏร่วมกันของตัวอักษรในชุดข้อมูลในการกำหนดขอบเขตของหน่วยคำย่อย

(3) sefr: การตัดแบ่งคำจาก Model machine learning จากบทความทางวิชาการชื่อ “Stacked Ensemble Filter and Refine for Word Segmentation” [10]



รูปที่ 4 ตัวอย่างข้อความที่ผ่านกระบวนการเตรียมข้อมูล (Dara Preprocessing)

3.3 Feature Extraction

เป็นกระบวนการแปลงคุณลักษณะต่างๆจากข้อความให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้ใน Model เช่น การแปลงข้อมูลที่เป็นข้อความให้เป็นชุดตัวเลข ในงานวิจัยนี้ ใช้กระบวนการที่เรียกว่า TF-IDF

TF-IDF หรือ Term Frequency-Inverse Document Frequency คือวิธีที่ไว้หาคำ หรือ Term ที่สำคัญจากเอกสาร (Document) โดยดูจากเนื้อหาโดยรวมทั้งหมด โดย TF-IDF เกิดจาก TF (Term Frequency) คือ ความถี่ของคำดังสมการที่ 1 และ IDF (Inverse Document Frequency) ใช้สำหรับวัดความสำคัญของคำๆนั้นเปรียบเทียบจากจำนวนการปรากฏของคำๆนั้นในเอกสารทั้งหมด โดยหากคำๆนั้นปรากฏขึ้นเป็นจำนวนมากในหลายๆเอกสาร ความสำคัญก็จะถูกลดลงไป ดังสมการที่ 2

$$\mathrm{TF}(\text{ของคำ}) = \frac{\text{จำนวนของคำนั้นๆ ในเอกสาร}}{\text{จำนวนของคำทั้งหมดในเอกสาร}} \qquad (1)$$

$$\mathrm{IDF}(\text{ของคำ}) = \log\left(\frac{\text{จำนวนเอกสารทั้งหมดที่ใช้พิจารณา}}{\text{จำนวนเอกสารที่มีคำคำนั้นปรากฏอยู่}}\right) \qquad (2)$$

จะได้ การหาค่า TF-IDF เพื่อหาคำที่มีความสำคัญ ตามสมการที่ 3

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF}\ \mathrm{x}\ \mathrm{IDF} \qquad (3)$$

3.4 การแบ่งข้อมูลเพื่อใช้สำหรับฝึก Model และทดสอบ Model (Train/Test Split)

3.3.1) Train Data Set

ข้อมูลที่เลือกใช้ในการ train Model เป็นข้อมูลจาก Wisesight-Sentiment ซึ่งเป็นข้อความภาษาไทยจากสื่อสังคมออนไลน์ พร้อมกับป้ายกำกับความรู้สึก (เชิงบวก, เป็นกลาง, เชิงลบ และคำถาม) ตามตารางที่ 1 รวม 26,737 ข้อความ โดยจะแบ่งข้อมูลตามสัดส่วน ดังนี้ ข้อมูลสำหรับ Train 76% ข้อมูลสำหรับ Test 10% และข้อมูลสำหรับ Validate 14%

ประเภทข้อความ	จำนวนข้อความ	ร้อยละ
เชิงบวก	4,778	0.18
เป็นกลาง	14,561	0.54
เชิงลบ	6,823	0.26
คำถาม	575	0.02
รวม	**26,737**	**1.00**

ตารางที่ 1 จำนวนข้อความและร้อยละของจำนวนข้อความ จากข้อมูล Wisesight Sentiment โดยแบ่งตามประเภทข้อความตามความรู้สึก

3.3.2) Test Data Set

ข้อมูลที่ใช้สำหรับการทดสอบประสิทธิภาพ Model มี 2 ชุด คือ ข้อมูลจาก Wisesight Sentiment ที่เป็น Validate set จำนวน 3,610 ข้อความที่มีสัดส่วนของความรู้สึกใกล้เคียงกับข้อมูลทั้งหมดในตารางที่ 1 และข้อมูลจาก Twitter เกี่ยวกับการเมืองที่รวบรวมมา โดยจะทำการสุ่มเลือกมาเพียง 1,000 ข้อความ พร้อมกับทำการกำกับประเภทข้อความตามความรู้สึก (เชิงบวก, เป็นกลาง, เชิงลบ และคำถาม) ตามตารางที่ 2 ของแต่ละข้อความด้วยการพิจารณาจากตัวผู้วิจัยเอง

ประเภทข้อความ	จำนวนข้อความ	ร้อยละ
เชิงบวก	137	0.14
เป็นกลาง	487	0.49
เชิงลบ	348	0.35
คำถาม	28	0.02
รวม	**1,000**	**1.00**

ตารางที่ 2 จำนวนข้อความและร้อยละของจำนวนข้อความ จากข้อมูล Tweet ที่สุ่มเลือกมาจำนวน 1,000 ข้อความ โดยแบ่งตามประเภทข้อความตามความรู้สึก

3.5 การวัดประสิทธิภาพ Model (Model Evaluation)

ผู้วิจัยได้ทำการเปรียบเทียบ 2 Models ระหว่าง WangchanBERTa กับ Logistic Regression เพื่อใช้ในการจำแนกประเภทข้อความตามความรู้สึก (เชิงบวก, เป็นกลาง, เชิงลบ และคำถาม) เพื่อคัดเลือก Model ที่มีประสิทธิภาพสูงสุดในการจำแนกความรู้สึกไปใช้ในกรอบการวิเคราะห์รูปแบบการลงคะแนนเสียงเลือกพรรคการเมืองในกลุ่มผู้ใช้ Twitter

3.5.1) WangchanBERTa

ในการวิเคราะห์ความรู้สึกของข้อความภาษาไทยได้มีการพัฒนาเทคนิตขึ้นมาอย่างต่อเนื่อง โดย Model ที่ถูกพัฒนาขึ้นมาคือ WangchanBERTa ซึ่งเป็นการ Train Language Model สถาปัตยกรรม BERT ด้วยข้อมูลภาษาไทยขนาดใหญ่ที่มีความหลากหลายและถูกทำความสะอาดมากที่สุด โดยใช้กฎการจัดการข้อมูลที่ถูกสร้างขึ้นมาเพื่อภาษาไทยโดยเฉพาะ ซึ่ง WangchanBERTa สามารถทำได้หลาย Task แต่ใน Task ที่ใช้ในการจำแนกความรู้สึกจะใช้ชุดข้อมูลของ Wisesight Sentiment ในการ Train ซึ่งเป็นข้อมูลความรู้สึกที่ได้มาจากสื่อสังคมออนไลน์ (Social Media)

สำหรับ WangchanBERTa นั้นถือได้ว่าเป็น Pretrain model ตัวหนึ่งที่ได้ถูก Train มาอย่างดีแล้วบนข้อมูลขนาดใหญ่ ดังนั้นในงานวิจัยนี้จึงแค่นำข้อมูลที่ผ่านการทำความสะอาด (Data Cleansing) และ Data Preprocessing เรียบร้อยแล้วนำเข้าสู่ Model เพื่อให้ Model จำแนกความรู้สึกของข้อความ จากนั้นทำการ Test เพื่อวัดประสิทธิภาพโมเดล

โดย Model WangchanBERTa สามารถแบ่งแยกย่อยได้อีก 3 Models ซึ่งแตกต่างกันที่วิธีการตัดคำ (Word Tokenized) และชุดข้อมูลที่ใช้ในการ train

(1) wangchanberta-base-att-spm-uncased:

เป็น Model ที่ถูก Train ด้วยชุดข้อมูลภาษาไทยที่มีขนาดใหญ่ที่สุด 78.5GB ประกอบด้วยข้อมูลจาก Social Media เช่น Facebook, Twitter, Pantip เป็นต้น และข้อมูลเปิด เช่น ข่าว, หนังสือ, สารานุกรม เป็นต้น โดยตัดคำแบบ spm คือตัดแบ่งหน่วยคำย่อย (subword-level tokenization) ด้วยไลบรารี่ SentencePiece

(2) wangchanberta-base-wiki-newmm:

เป็น Model ที่ถูก Train ด้วยชุดข้อมูลภาษาไทยที่มาจาก Wikipedia โดยตัดคำแบบ newmm เป็นการตัดแบ่งคำจาก dictionary ของคำในภาษาไทยด้วย maximal matching algorithm ด้วยไลบรารี่ PyThaiNLP

(3) wangchanberta-base-wiki-sefr:

เป็น Model ที่ถูก Train ด้วยชุดข้อมูลภาษาไทยที่มาจาก Wikipedia โดยตัดคำแบบ sefr เป็นการตัดแบ่งคำจาก Model machine learning

3.5.2) Logistic Regression

Logistic Regression หรือ Maximum Entropy Model นับว่าเป็น machine learning model ที่สำคัญ เพราะไม่จำเป็นต้อง finetune มากและสามารถพัฒนามาประยุกต์ใช้ได้อย่างรวดเร็ว และมีประสิทธิภาพ

โดยในงานวิจัยนี้ จะ Train Model โดยใช้ข้อมูลจาก Wisesight-sentiment ที่เป็น Train set จำนวน 20,320 ข้อความ

Dataset		Tweet Test Set (1,000 messages)				Wisesight Validate Set (3,610 messages)			
No	Model	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
1	wangchanberta-base-att-spm-uncased	0.76	0.76	0.76	0.74	0.84	0.84	0.84	0.84
2	wangchanberta-base-wiki-newmm	0.62	0.65	0.62	0.59	0.81	0.81	0.81	0.8
3	wangchanberta-base-wiki-sefr	0.49	0.24	0.49	0.32	0.54	0.38	0.54	0.38
4	Logistic Regression	0.59	0.57	0.59	0.56	0.73	0.73	0.73	0.71

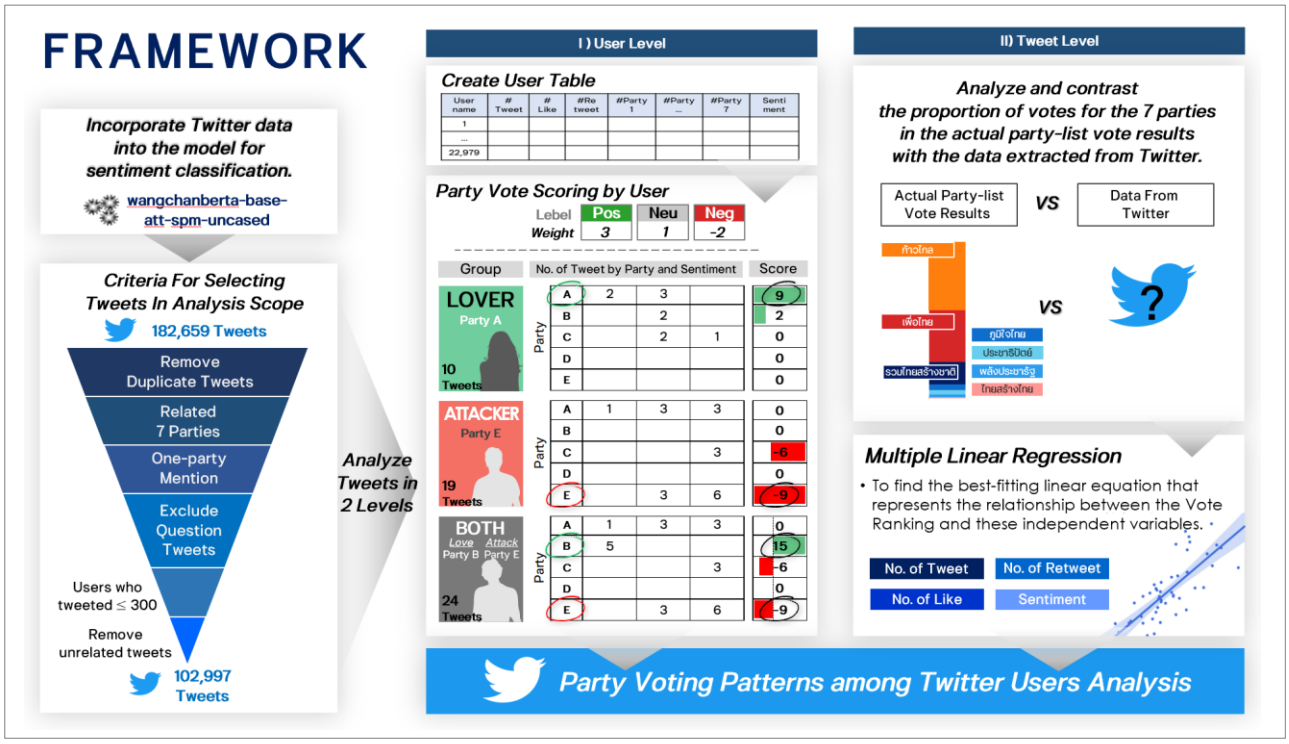
ตารางที่ 3 เปรียบเทียบประสิทธิภาพ Model บนข้อมูล 2 ชุด

จากตารางที่ 3 ตารางเปรียบเทียบประสิทธิภาพ Model บนข้อมูล 2 ชุด พบว่า Model wangchanberta-base-att-spm-uncased แสดงประสิทธิภาพสูงสุดในข้อมูลทั้งสองชุด เมื่อเปรียบเทียบกับอีก 3 Models โดย wangchanberta-base-att-spm-uncased มีความแม่นยำ (Accuracy) สูงสุดที่ 76% บนข้อมูลชุด Tweet ที่สุ่มเลือกมาจาก Twitter และความแม่นยำ (Accuracy) สูงสุดที่ 84% บนข้อมูลชุด Wisesight Validate Set รองลงมาเป็น Model wangchanberta-base-wiki-newmm, Logistic Regression และ wangchanberta-base-wiki-sefr ตามลำดับ

จากผลการเปรียบเทียบดังกล่าว ในงานวิจัยนี้จึงเลือกใช้ผลการทำนายจาก Model wangchanberta-base-att-spm-uncased ในการจำแนกความรู้สึกของข้อความใน Twitter เพื่อใช้ในกรอบการวิเคราะห์ (Framework) ต่อไป

3.6 กรอบการวิเคราะห์รูปแบบการลงคะแนนเสียงเลือกพรรคการเมืองในกลุ่มผู้ใช้ Twitter

เมื่อได้ Model wangchanberta-base-att-spm-uncased ที่จะใช้ในการจำแนกความรู้สึกของข้อความใน Twitter แล้ว ขั้นตอนต่อไป คือ ขั้นตอนการคัดเลือกข้อความที่เกี่ยวข้อง จากนั้นจะนำข้อความทั้งหมดไปวิเคราะห์ต่อใน 2 ระดับ ระดับแรกเป็นการวิเคราะห์ในระดับผู้ใช้ Twitter ว่าผู้ใช้ Twitter ชื่นชอบหรือโจมตีพรรคการเมืองใด หรือมีทั้งพรรคที่ชอบและพรรคที่ไม่ชอบ จากนั้นจึงวิเคราะห์ในระดับข้อความ โดยทำการเปรียบเทียบสัดส่วนของการกล่าวถึงพรรคการเมืองในข้อมูล Twitter ทั้งหมดกับข้อมูลผลการเลือกตั้งแบบบัญชีรายชื่อที่เกิดขึ้นจริงว่ามีความสอดคล้องกันอย่างไรบ้าง จากนั้นใช้ Multiple Linear Regression Model เข้ามาช่วยพิจารณาว่าปัจจัยใดที่มีผลต่อลำดับการเลือกตั้งพรรคการเมืองแบบบัญชายชื่อ สุดท้ายก็จะนำไปสู่บทสรุปของข้อมูลเชิงลึกที่สามารถนำไปใช้ต่อได้ ตามรูปที่ 5



รูปที่ 5 กรอบการวิเคราะห์รูปแบบการลงคะแนนเสียงเลือกพรรคการเมืองในกลุ่มผู้ใช้ Twitter

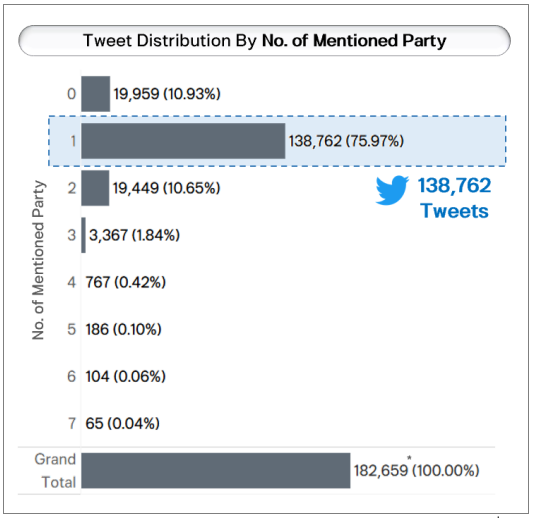
3.6.1) การคัดเลือกข้อความที่เกี่ยวข้อง (Criteria for Selecting Tweets in Analysis Scope)

จากข้อความ Tweet ทั้งหมดที่รวบรวมมาได้ 182,659 ข้อความ จะต้องทำการคัดเลือก Tweet ที่เกี่ยวข้องผ่านเงื่อนไขต่างๆ 6 ข้อ ดังนี้

(1) ลบข้อความ (Tweet) ที่มีรหัสของข้อความซ้ำกัน (Twee ID duplicate) ซึ่งอาจเกิดขึ้นในขั้นตอนของการรวบรวมข้อมูล

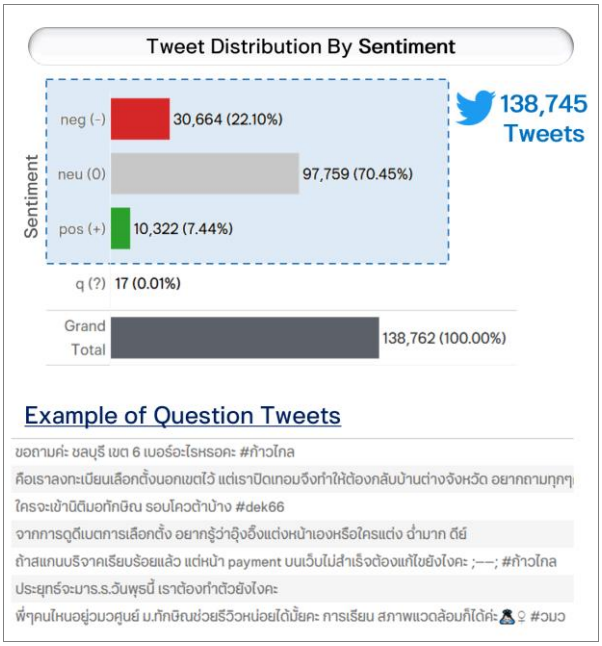
(2) เลือกเฉพาะข้อความ (Tweet) ที่เกี่ยวข้องกับพรรคการเมือง 7 พรรคหลักที่สนใจ หากข้อความนั้นพูดถึงพรรคการเมืองอื่นๆนอกจาก 7 พรรรคที่สนใจก็จะถูกคัดออก

(3) เลือกเฉพาะข้อความ (Tweet) ที่กล่าวถึงพรรคการเมืองเพียงพรรคเดียวในแต่ละข้อความ เพื่อให้ง่ายในการวิเคราะห์ความรู้สึก เพราะถ้าในหนึ่งข้อความมีการกล่าวถึงพรรคการเมืองมากกว่าหนึ่งพรรค อาจจะเป็นข้อความที่ชื่นชมพรรคการเมือง ก. แต่โจมตีพรรคการเมือง ข. ก็เป็นไปได้ และเมื่อพิจารณาความรู้สึกของข้อความร่วมด้วย ก็อาจจะทำให้เกิดความกำกวมได้ ดังนั้นจึงเลือกเฉพาะข้อความ (Tweet) ที่กล่าวถึงพรรคการเมืองเพียงพรรคเดียว ตามรูปที่ 6



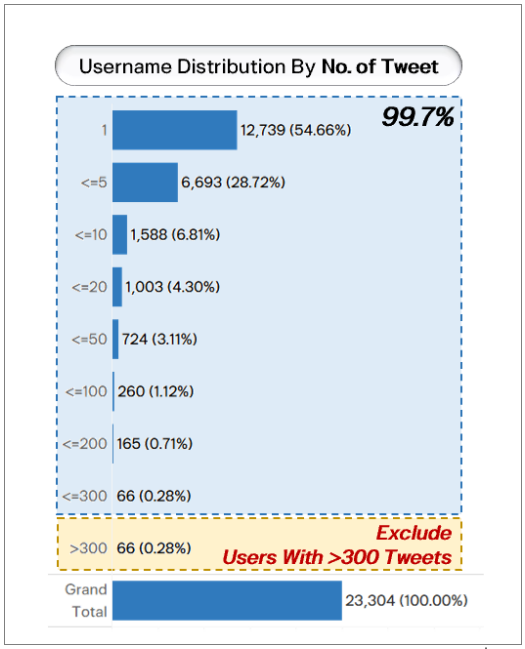
รูปที่ 6 การกระจายตัวของข้อความ (Tweet) โดยแบ่งตามจำนวนพรรคการเมืองที่กล่าวถึงในแต่ละข้อความ

(4) เลือกเฉพาะข้อความ (Tweet) ที่ Modelทำนายความรู้สึกออกมาว่าเป็นเชิงบวก, เป็นกลาง และเชิงลบ โดยคัดข้อความที่เป็นคำถามออก เพราะไม่สามารถวิเคราะห์ข้อความคำถามได้ และอาจเป็นข้อความที่ไม่เกี่ยวข้องตามกรอบการวิเคราะห์ ตามรูปตัวอย่างรูปที่ 7

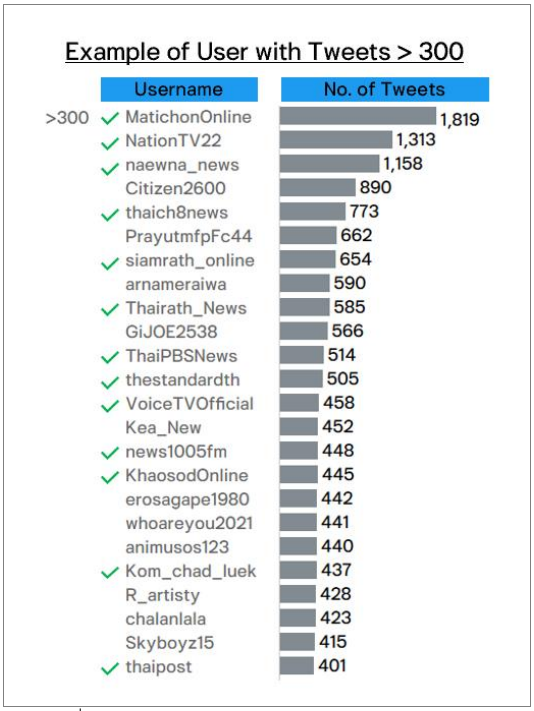


รูปที่ 7 การกระจายตัวของข้อความ (Tweet) โดยแบ่งตามประเภทข้อความตามความรู้สึก และตัวอย่างข้อความประโยคคำถามที่ไม่เกี่ยวข้อง

(5) เลือกเฉพาะข้อความ (Tweet) ที่ถูกโพสต์โดยผู้ใช้ที่โพสต์ข้อความน้อยกว่า 300 ข้อความในช่วงเวลาที่สนใจ (100 วัน) เนื่องจากเมื่อพิจารณาผู้ใช้ที่โพสต์ข้อความที่มีจำนวนมากกว่า 300 ข้อความ ส่วนใหญ่เป็นผู้ใช้ที่เป็นสำนักข่าว ผู้วิจัยจึงเลือกที่จะตัดออก เพราะอยากจะพิจารณาเฉพาะผู้ใช้ Twitter ทั่วไปเท่านั้น ดังรูปที่ 8 และรูปที่ 9



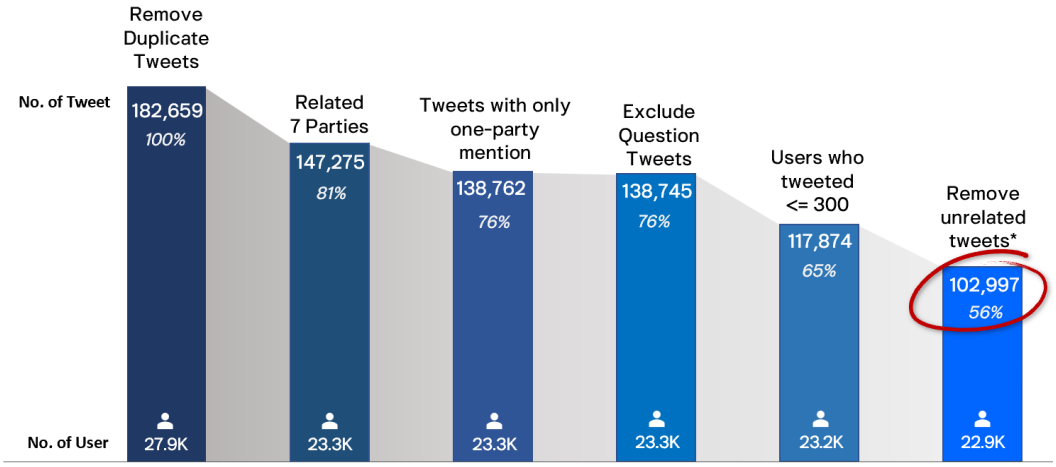
รูปที่ 8 การกระจายตัวของผู้ใช้ Twitter โดยแบ่งตามจำนวนข้อความ (No. of Tweet) ที่โพสต์ในช่วงเวลาที่สนใจ (100 วัน)



รูปที่ 9 ตัวอย่างรายชื่อผู้ใช้ (Username) และจำนวนข้อความ (No. of Tweet) ที่โพสต์

(6) คัดข้อความที่ไม่เกี่ยวข้องออก เช่น ข้อความที่เป็นการโฆษณาอื่นๆที่ไม่เกี่ยวข้องกับการแสดงความคิดเห็นทางการเมือง แต่มีการติด tag ตามคำค้นหาสำคัญ (Keyword) ทำให้ข้อความเหล่านี้ถูกรวบรวมเข้ามาด้วย โดยผู้วิจัยจะทำการลบข้อความ (Tweet) ที่ประกอบไปด้วยคำเหล่านี้ออกไปจากกรอบการวิเคราะห์ *“ช้อปเลย|โปรดีๆ|ส่วนลด|งานออนไลน์|พร้อมส่ง|ป้ายยา|ช้อป|รับหิ้ว|รวมส่ง|โค้ดลด|แถมฟรี|ลดราคาเหลือ|สนใจทัก|ไม่มีขั้นต่ำ|ส่งฟรี|พร้อมโอน|ปล่อยกู้รายเดือน|ราคาชิ้นละ|ส่งต่อเสื้อผ้า|เสื้อเปิดไหล่|ดีลเดือด|ราคาถูก|ดูดวง|พร้อมส่ง|Spotify|ถอนได้ไม่อั้น|ซื้อตู้เย็น|เสื้อกล้าม|มีงานมาแนะนำ|เกมมือถือ|เสื้อสงกรานต์|แอนจักรพงษ์|เที่ยวงาน|สอบถามจองคิว|ยำขนมจีน|ครีมบำรุงหน้า|มอทักษิณ,|Google Search Trend|นค.ทักษิณ|รับสมัครงาน”*

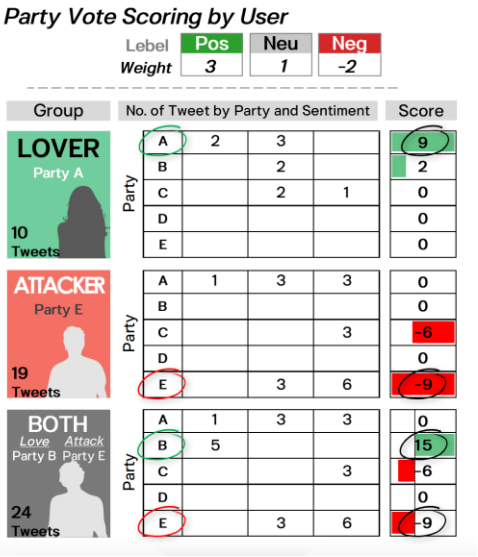
จากข้อความ (Tweet) ทั้งหมด 182,659 ข้อความ เมื่อคัดกรองผ่านเงื่อนไขทั้ง 6 ข้อที่กล่าวมาข้างต้น ทำให้เหลือข้อความที่จะนำไปวิเคราะห์ต่อในขั้นตอนต่อไปทั้งสิ้น 102,997 ข้อความ ตามรูปที่ 10



รูปที่ 10 เงื่อนไขการคัดเลือกข้อความที่เกี่ยวข้อง

3.6.2) การวิเคราะห์ในระดับผู้ใช้ Twitter (User Level)

เนื่องจากข้อมูลที่เก็บรวบรวมมาเป็นข้อมูลรายข้อความ (Tweet) จึงต้องทำการสร้างตารางผู้ใช้ขึ้นมา โดยตารางจะประกอบไปด้วยชื่อผู้ใช้, จำนวนข้อความ (No. of Tweet) ที่โพสต์, จำนวนการได้รับการกดถูกใจ (No. of Like), จำนวนการถูกแชร์ต่อ (No. of Retweet) โดยแยกตามพรรคการเมือง และแยกตามประเภทความรู้สึก ซึ่งจากข้อความทั้งหมด 102,997 ข้อความ เป็นข้อความที่ถูกโพสต์จากผู้ใช้ทั้งสิ้น 22,979 คน (Username)



รูปที่ 11 ตัวอย่างการคำนวณคะแนนความนิยมพรรคการเมือง โดยแบ่งเป็น 3 กลุ่ม ได้แก่ Lover, Attacker และ Both

เมื่อได้ตารางผู้ใช้แล้ว จึงนำข้อมูลที่ได้ไปคำนวณคะแนนตามเกณฑ์ที่กำหนด ดังรูปที่ 11 โดยกำหนดให้ข้อความที่เป็นกลางมีน้ำหนักเท่ากับ 1 เนื่องจากการที่พรรคการเมืองนั้นถูกกล่าวถึงโดยผู้ใช้ พรรคการเมืองนั้นก็น่าจะมีแนวโน้มที่จะเป็นที่สนใจของผู้ใช้ สำหรับข้อความเชิงบวก มีน้ำหนักเท่ากับ 3 เพราะผู้ใช้น่าจะชื่นชอบพรรคการเมืองนั้นเป็นพิเศษ ขณะที่ข้อความเชิงลบ มีน้ำหนักเท่ากับ -2 เนื่องจากพรคการเมืองนั้นน่าจะเป็นพรรคการเมืองที่ผู้ใช้โจมตีหรือไม่ชอบ จากนั้นก็จะนำข้อมูลผู้ใช้มาดูรายคน เพื่อคำนวณคะแนนของแต่ละพรรคการเมืองรายบุคคล โดยนำจำนวน Tweet คูณด้วยน้ำหนักตามประเภทความรู้สึก แล้วบวกรวมกันเป็นคะแนนรายพรรคการเมือง

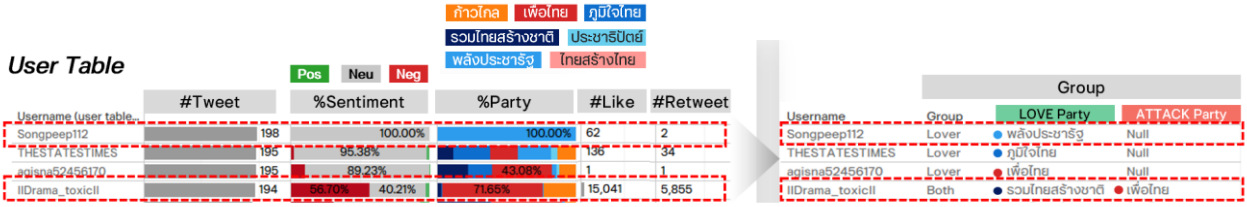
ยกตัวอย่างการคำนวณ เช่น นางสาว ก. (กลุ่ม Lover) โพสต์ข้อความทั้งหมด 10 ข้อความ โดยกล่าวถึงพรรค A ทั้งหมด 5 ข้อความ แบ่งเป็น เชิงบวก 2 ข้อความและเป็นกลาง 3 ข้อความ ดังนั้นสำหรับนางสาว ก. พรรคการเมือง A มีคะแนนเท่ากับ (เชิงบวก 2 x 3) + (เป็นกลาง 3 x 1) = 6+3 = 9 คะแนน ส่วนพรรค B จะได้คะแนนเท่ากับ (เป็นกลาง 2 x 1) = 2 คะแนน และพรรค C ได้คะแนนเท่ากับ (เป็นกลาง 2 x 1) + (เชิงลบ 1 x (-2)) = 2+ (-2) = 0 คะแนน

ดังนั้นสำหรับนางสาว ก. จะมีคะแนนแต่ละพรรค ดังนี้ พรรค A: 9 คะแนน, พรรค B: 2 คะแนน และพรรค C: 0 คะแนน จะเห็นว่าไม่มีพรรคไหนที่คะแนนติดลบเลย จึงสรุปได้ว่านางสาว ก. เป็นกลุ่ม Lover หรือกลุ่มผู้ชื่นชอบ โดยชอบพรรค A มากที่สุด เพราะพรรค A มีคะแนนสูงสุด

ส่วนนาย ข. (กลุ่ม Attacker) จะเห็นว่าไม่มีคะแนนพรรคการเมืองไหนเลยที่มีค่าเป็นบวก ดังนั้นนาย ข. คือกลุ่ม Attacker หรือกลุ่มผู้โจมตี โดยพรรคที่คะแนนติดลบสูงสุดคือพรรค E จึงสรุปได้ว่า นาย ข. อยู่ในกลุ่มที่โจมตีพรรค E

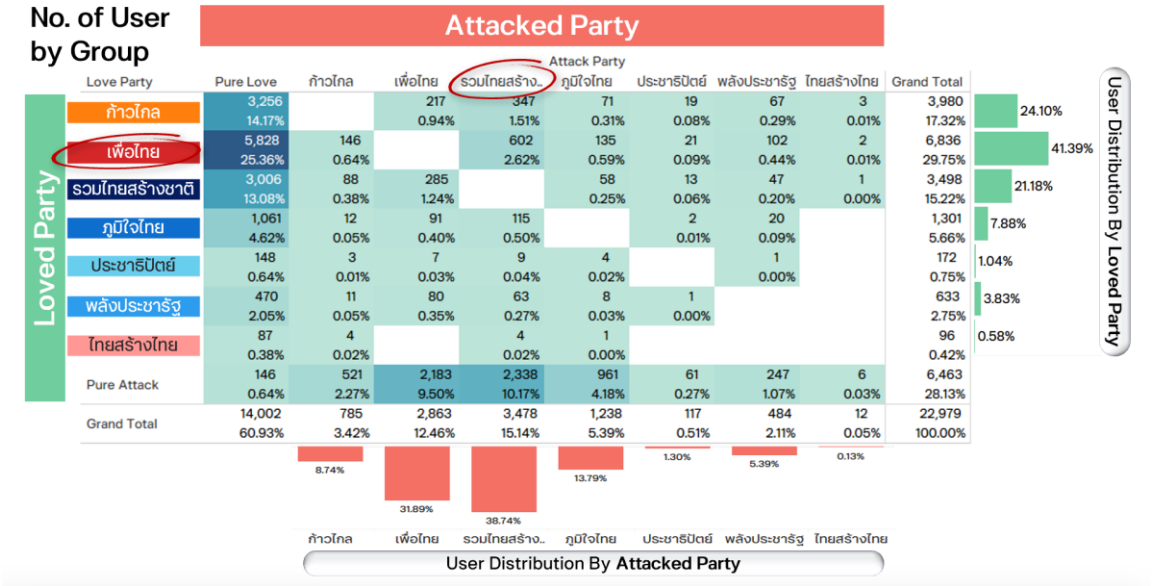
ส่วนนาย ค. (กลุ่ม Both) มีคะแนนพรรคการเมืองทั้งที่เป็นบวกและลบ จึงสรุปได้ว่า นาย ค. อยู่กลุ่ม Both มีทั้งพรรคที่ชอบคือพรรค B และพรรคที่ไม่ชอบ คือพรรค E

จากวิธีการคำนวณดังกล่าว ทำให้ผู้วิจัยสามารถอนุมานได้ว่าผู้ใช้แต่ละคนชื่นชอบหรือโจมตีพรรคการเมืองใด



รูปที่ 12 ตัวอย่างตารางข้อมูลผู้ใช้ Twitter

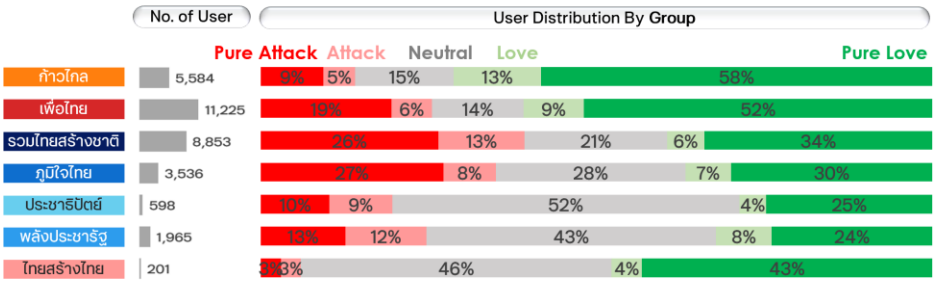
จากรูปที่ 12 ผู้ใช้ชื่อ 'Songpeep112' โพสต์ข้อความทั้งหมด 198 ข้อความ (Tweet) โดยทุกข้อความเป็นความรู้สึกกลางๆ ที่พูดถึงพรรคพลังประชารัฐทั้งหมด เมื่อผ่านการคำนวณคะแนนความนิยมแล้ว จะสรุปได้ว่าผู้ใช้รายนี้เป็นกลุ่ม Lover ที่ชื่นชอบพรรคพลังประชารัฐ หรือผู้ใช้ที่ใช้ชื่อว่า 'IIDrama_toxicII' ได้มีการโพสต์ข้อความทั้งสิ้น 194 ข้อความ (Tweet) โดยแบ่งเป็นข้อความเชิงลบ 57%, ข้อความที่เป็นกลาง 40% และข้อความเชิงบวก 3% และกล่าวถึงพรรคเพื่อไทยมากที่สุด 72% และพรรคอื่นๆรองลงมา หากพิจารณาข้อมูลเพียงเท่านี้ อาจจะสรุปได้ว่าผู้ใช้รายนี้ชื่นชอบพรรคเพื่อไทยเพราะมีการกล่าวถึงพรรคเพื่อไทยเป็นจำนวนมาก แต่การสรุปแบบนี้ก็อาจจะไม่ถูกต้อง เพราะการที่กล่าวถึงเป็นจำนวนมาก อาจจะเป็นข้อความเชิงลบที่โจมตีพรรคเพื่อไทยก็ได้ ดังนั้นจึงต้องมีการคำนวณคะแนนความนิยมพรรคก่อน ซึ่งผลออกมาปรากฏว่าผู้ใช้รายนี้เป็นกลุ่ม Both คือมีพรรคที่ชื่นชอบ นั่นก็คือพรรครวมไทยสร้างชาติ และโจมตีพรรคเพื่อไทย



รูปที่ 13 จำนวนผู้ใช้ Twitter และสัดส่วนของผู้ใช้ แบ่งตามกลุ่มพรรคที่ชื่นชอบและพรรคที่โจมตี

จากผู้ใช้ทั้งหมด 22,979 คน สามารถแบ่งกลุ่มได้ 3 กลุ่ม ดังนี้ 1) Lover 13.9 พันคน (60%), 2) Attacker 6.5 พันคน (28%) และ 3) Both 2.7 พันคน (12%)

จากรูปที่ 13 พรรคที่ผู้ใช้ชื่นชอบมากที่สุด คือ พรรคเพื่อไทย คิดเป็น 41.39% ของกลุ่ม Lover ทั้งหมด ขณะที่พรรครวมไทยสร้างชาติเป็นพรรคที่ผู้ใช้โจมตีมากที่สุด คิดเป็น 38.74% ของกลุ่ม Attacker ทั้งหมด



รูปที่ 14 จำนวนผู้ใช้ Twitter ที่กล่าวถึงแต่ละพรรค และสัดส่วนของกลุ่มผู้ใช้

จากรูปที่ 14 พรรคที่มีผู้ใช้กล่าวถึงมากที่สุด จำนวน 11,225 คน คือ พรรคเพื่อไทย รองลงมาเป็นพรรคก้าวไกล และพรรครวมไทยสร้างชาติ ตามลำดับ ต่อมาพิจารณาสัดส่วนของกลุ่มผู้ใช้ โดยแบ่งเป็น 5 กลุ่ม ดังนี้

(1) Pure Attack คือ กลุ่มที่โจมตีพรรคการเมืองนั้นๆเพียงแค่พรรคเดียวเท่านั้น โดยที่ไม่สนใจพรรคการเมืองอื่นๆเลย

(2) Attack คือ กลุ่มที่โจมตีพรรคการเมืองนั้นๆ และมีพรรคอื่นๆที่ชื่นชอบอยู่แล้ว

(3) Neutral คือ กลุ่มที่กล่าวถึงพรรคนั้นๆแบบกลางๆ

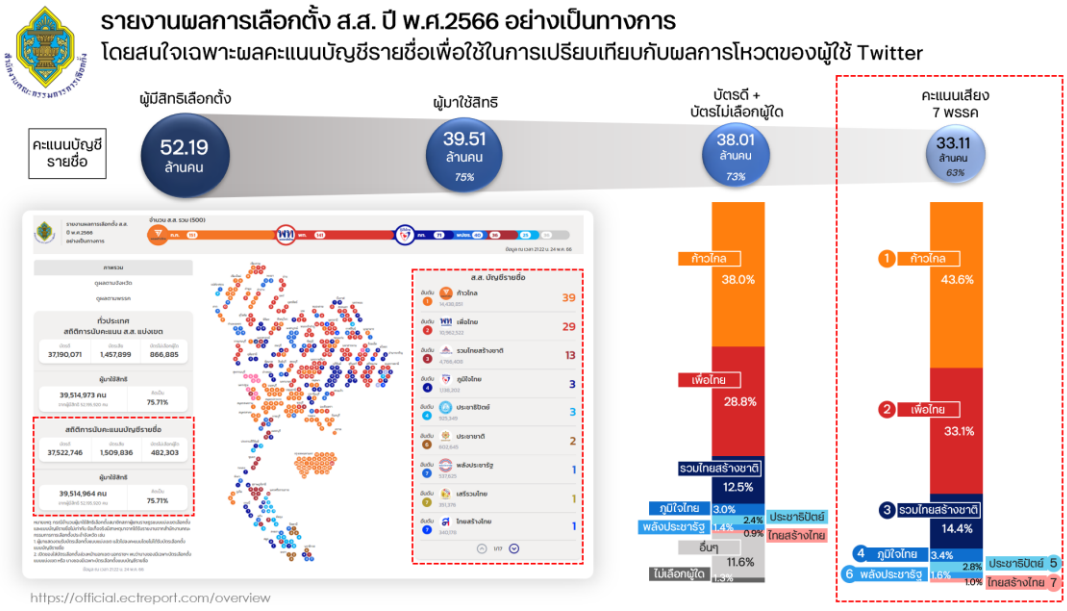
(4) Love คือกลุ่มที่ชื่นชอบพรรคการเมืองนั้นๆ พร้อมๆกับการโจมตีพรรคการเมืองอื่นๆไปด้วย

(5) Pure Love คือ กลุ่มที่ชื่นชอบพรรคการเมืองนั้นๆ โดยที่ไม่สนใจพรรคการเมืองอื่นๆเลย

จากรูปที่ 14 จะเห็นว่า 71% (Pure Love + Love) ของคนที่กล่าวถึงพรรคก้าวไกล ล้วนเป็นคนที่ชื่นชอบพรรคก้าวไกล ขณะที่เกือบ 40% (Pure Attack + Attack) ของคนที่กล่าวถึงพรรครวมไทยสร้างชาติเป็นกลุ่มคนที่โจมตีพรรค

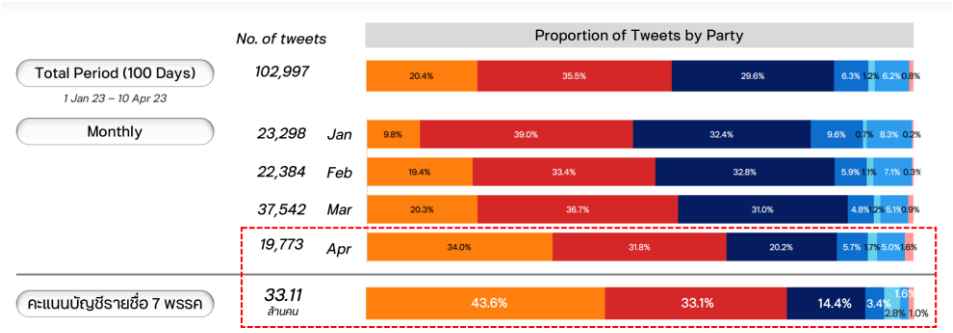
3.6.3) การวิเคราะห์ในระดับข้อความ (Tweet Level)

ในงานวิจัยนี้จะพิจารณาผลคะแนนการเลือก ส.ส. ปี พ.ศ.2566 แบบบัญชีรายชื่อ [11] เท่านั้น และสนใจเฉพาะคะแนนเสียงของ 7 พรรคการเมืองที่สนใจเท่านั้น โดยมีผู้มาใช้สิทธิ์เลือก 7 พรรคการเมืองนี้ จำนวน 33.11 ล้านคน จากรูปที่ 15 จะเห็นว่าพรรคที่ได้คะแนนเป็นอันดับ 1 คือพรรคก้าวไกล 43.6%, อันดับ 2 พรรคเพื่อไทย 33.1%, อันดับ 3 พรรครวมไทยสร้างชาติ 14.4%, อันดับ 4 พรรคภูมิใจไทย 3.4%, อันดับ 5 พรรคประชาธิปัตย์ 2.8%, อันดับ 6 พรรคพลังประชารัฐ 1.6% และอันดับ 7 พรรคไทยสร้างไทย 1.0%



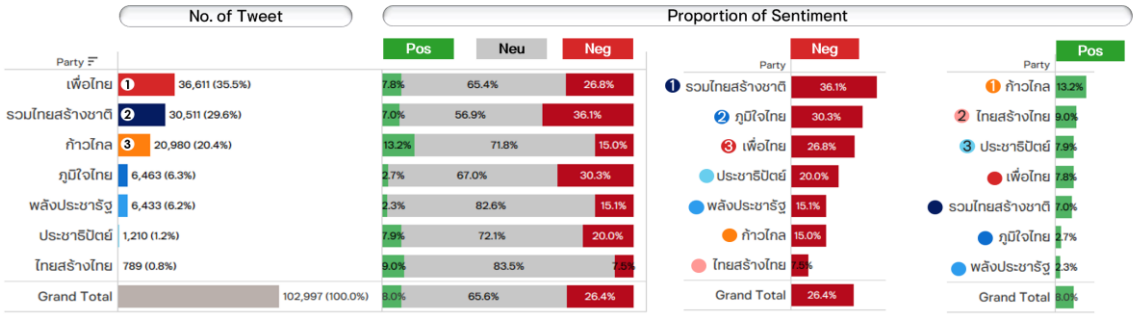
รูปที่ 15 สัดส่วนผลคะแนนการเลือกตั้งแบบบัญชีรายชื่อของ 7 พรรคการเมือง

ตามรูปที่ 16 หากพิจารณาแค่จำนวนข้อความ (No. of Tweet) เท่านั้น ในระยะเวลา 100 วัน (1 ม.ค. 2566 – 10 เม.ย. 2566) สัดส่วนของพรรคเพื่อไทยสูงสุดที่ 35.5% รองลงมาเป็นพรรครวมไทยสร้างชาติ 29.6% และพรรคก้าวไกล 20.4% ซึ่งอันดับคะแนนไม่สอดคล้องกับผลการเลือกตั้งจริงที่เกิดขึ้น แต่ถ้าพิจารณาเป็นรายเดือนยิ่งระยะเวลาเข้าใกล้วันเลือกตั้งมากเท่าไหร่ สัดส่วนของอันดับพรรคการเมืองจะเริ่มใกล้เคียงกับสัดส่วนผลคะแนนจริงมากขึ้น โดยภายใน 10 วันแรกของเดือนเมษายน พ.ศ.2566 มีจำนวนข้อความ (No. of Tweet) ทั้งสิ้น 19,773 ข้อความ และมีสัดส่วนของพรรคก้าวไกลมากที่สุดเป็นอันดับหนึ่งที่ 34% รองลงมาเป็นเพื่อไทย 31.8%, พรรครวมไทยสร้างชาติ 20.2% และพรรคอันดับสี่คือ พรรคภูมิใจไทย 5.7% ซึ่งอันดับสัดส่วนคะแนน 4 พรรคแรกสอดคล้องกับสัดส่วนของผลคะแนนการเลือกตั้งจริง ดังนั้นจึงอนุมานได้ว่า ยิ่งข้อมูลใกล้วันเลือกตั้ง สัดส่วนคะแนนก็จะยิ่งใกล้เคียงกับผลคะแนนการเลือกตั้งที่เกิดขึ้นจริง



รูปที่ 16 จำนวนข้อความ (No. of Tweet) และสัดส่วนของข้อความในแต่ละพรรคการเมืองที่ถูกกล่าวถึง แบ่งตามระยะเวลา 100 วัน, รายเดือน โดยเปรียบเทียบกับสัดส่วนผลคะแนนเลือกตั้งแบบบัญชีรายชื่อ 7 พรรค

จากรูปที่ 17 หากพิจารณาแค่จำนวนข้อความ (No. of Tweet) พรรคเพื่อไทยจะมาเป็นอันดับ 1 ที่ 35.5% รองลงมาเป็นพรรครวมไทยสร้างชาติ และพรรคก้าวไกล ซึ่งอันดับคะแนนไม่สอดคล้องกับผลการเลือกตั้งจริงที่เกิดขึ้น ดังนั้นควรจะพิจารณาความรู้สึกของข้อความร่วมด้วย เมื่อเรียงลำดับสัดส่วนของข้อความที่เป็นเชิงลบจากมากไปน้อย จะพบว่าพรรครวมไทยสร้างชาติขึ้นมาเป็นอันดับ1 รองลงมาเป็นพรรคภูมิใจไทย และพรรคเพื่อไทย และเมื่อเรียงลำดับสัดส่วนของข้อความที่เป็นเชิงบวกจากมากไปน้อย พบว่าพรรคก้าวไกลขึ้นมาอันดับ 1 รองลงมาเป็นพรรคไทยสร้างไทย และพรรคประชาธิปัตย์

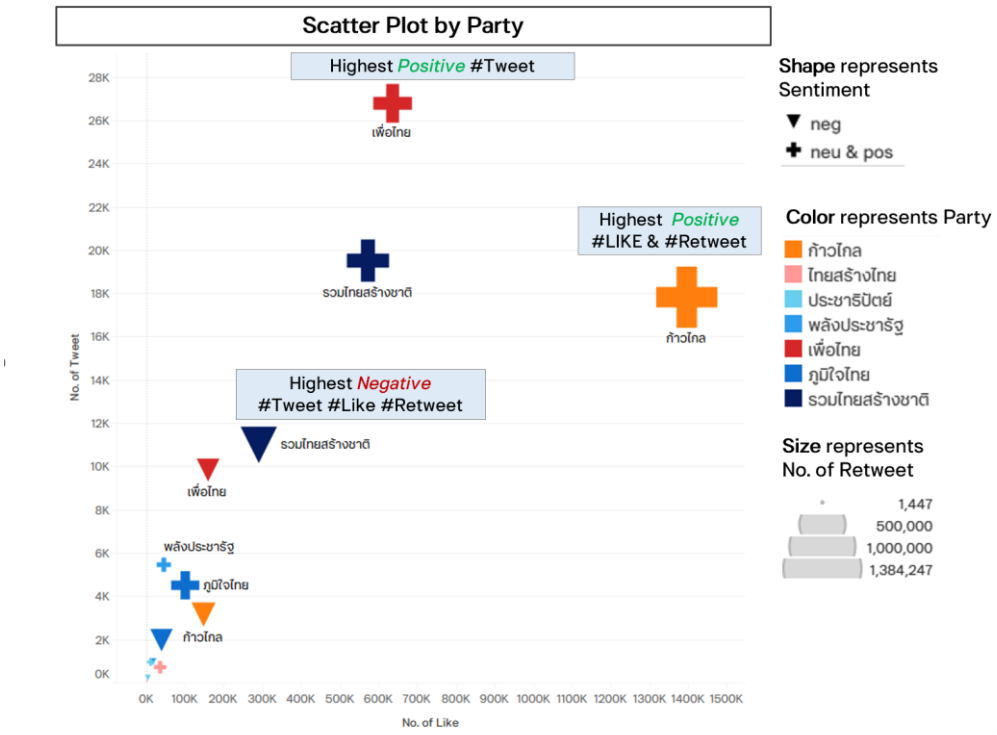


รูปที่ 17 จำนวนข้อความ (No. of Tweet) ในแต่ละพรรคการเมือง และสัดส่วนของข้อความแบ่งตามประเภทของความรู้สึก

เมื่อพิจารณาหลายๆปัจจัยรวมกันในหนึ่งกราฟเป็น Scatter plot ในรูปที่ 18 โดยกำหนดให้แกนตั้งเป็นจำนวนของข้อความ (No. of Tweet), แกนนอนเป็นจำนวนการกดถูกใจ (No. of Like), ขนาดของรูปเป็นจำนวนการแชร์ (No. of Retweet), สีแทนสีประจำพรรคการเมือง และรูปทรงแทนข้อความเชิงลบ และข้อความเชิงบวกกับกลางๆ จากรูปจะพบว่า

- พรรคเพื่อไทยมีจำนวนข้อความเชิงบวกและกลางๆสูงที่สุด (No. of Positive/Neutral Tweet)
- พรรคก้าวไกลมีจำนวนการกดถูกใจและการแชร์ข้อความในเชิงบวกและกลางๆสูงที่สุด (No. of Positive/Neutral Like and Retweet)

- พรรครวมไทยสร้างชาติมีจำนวนข้อความ จำนวนการกดถูกใจและการแชร์ในข้อความเชิงลบมากที่สุด (No. of negative tweet and Like and Retweet)



รูปที่ 18 Scatter Plot

3.6.4) Multiple Linear Regression

Multiple Linear Regression (MLR) เป็นสมการเชิงเส้นที่มีตัวแปรอิสระ (independent variables) หรือตัวแปร X มากกว่า 1 ตัวมาเป็นตัวกำหนดตัวแปรตาม (Dependent variable) หรือค่า Y ความหมายคือมีหลายปัจจัยที่ส่งผลกระทบต่อค่าที่ให้ความสนใจ โดย MLR เป็นเครื่องมือทางสถิติ โดยมีเป้าหมายเพื่อหาสมการที่สามารถอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม โดยการประมาณค่าสัมประสิทธิ์ (Coefficients) ที่ทำให้ Mean Squared Error (MSE) มีค่าน้อยที่สุด ดังสมการ (4)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \qquad (4)$$

โดยที่ Y แทนค่า ตัวแปรตาม (Dependent variable)

X_i แทนค่า ตัวแปรอิสระ (independent variables)

B_i แทนค่า ค่าสัมประสิทธิ์ (Coefficients)

ε แทนค่า ค่าความผิดพลาด (Error term or Residual)

เนื่องจากเป้าหมายของงานวิจัยในครั้งนี้ คือ ต้องการศึกษาว่าข้อมูลจาก Twitter แบบใดที่สามารถอธิบายหรือทำนายผลการเลือกตั้งได้ ดังนั้น สิ่งที่ผู้วิจัยสนใจก็คือ อันดับของผลคะแนนการเลือกตั้งแบบบัญชีรายชื่อของพรรคการเมือง 7 พรรคที่สนใจ ซึ่งเปรียบเสมือนกับตัวแปรตาม (Dependent variable) หรือค่า Y และข้อมูลจาก Twitter ที่เปรียบเสมือนตัวแปรอิสระ (independent variables) หรือตัวแปร X เช่น จำนวนข้อความ (No. of Tweet), จำนวนการได้รับการกดถูกใจ (No. of Like), จำนวนการถูกแชร์ต่อ (No. of Retweet) ที่ถูกแบ่งตามประเภทความรู้สึก โดยประยุกต์ใช้ Multiple Linear Regression เพื่อตอบคำถามดังกล่าว

			Independent Variables														
	Dependent Variable		All Sentiment			Pos			Neu			Neg			Pos + Neutral		
	Party	Score Rank	#Tweet	#Like	#Retweet	#Tweet_ Pos	#Like_ Pos	#Retweet_ Pos	#Tweet_ Neu	#Like_ Neu	#Retweet_ Neu	#Tweet_ Neg	#Like_ Neg	#Retweet_ Neg	#Tweet_ Pos+Neu	#Like_ Pos+Neu	#Retweet_ Pos+Neu
Actual Number	ก้าวไกล	1	20,980	1,544,421	1,609,575	2,767	255,837	242,641	15,060	1,139,891	1,141,606	3,153	148,693	225,328	17,827	1,395,728	1,384,247
	เพื่อไทย	2	36,611	797,018	755,372	2,845	82,358	49,519	23,942	554,321	503,941	9,824	160,339	201,912	26,787	636,679	553,460
	รวมไทยสร้างชาติ	3	30,511	865,664	1,170,978	2,150	95,019	109,344	17,351	478,687	549,930	11,010	291,958	511,704	19,501	573,706	659,274
	ภูมิใจไทย	4	6,463	140,938	488,411	177	13,790	44,180	4,329	86,436	249,370	1,957	40,712	194,861	4,506	100,226	293,550
	ประชาธิปัตย์	5	1,210	14,309	34,299	95	550	81	873	9,780	21,094	242	3,979	13,124	968	10,330	21,175
	พลังประชารัฐ	6	6,433	62,218	93,827	148	244	187	5,313	45,554	73,989	972	16,420	19,651	5,461	45,798	74,176
	ไทยสร้างไทย	7	789	37,659	58,568	71	13,462	24,157	659	23,025	32,964	59	1,172	1,447	730	36,487	57,121
		Total	102,997	3,462,227	4,211,030	8,253	461,260	470,109	67,527	2,337,694	2,572,894	27,217	663,273	1,168,027	75,780	2,798,954	3,043,003
% of Total	ก้าวไกล	1	0.2037	0.4461	0.3822	0.3353	0.5546	0.5161	0.2230	0.4876	0.4437	0.1158	0.2242	0.1929	0.2352	0.4987	0.4549
	เพื่อไทย	2	0.3555	0.2302	0.1794	0.3447	0.1786	0.1053	0.3546	0.2371	0.1959	0.3610	0.2417	0.1729	0.3535	0.2275	0.1819
	รวมไทยสร้างชาติ	3	0.2962	0.2500	0.2781	0.2605	0.2060	0.2326	0.2569	0.2048	0.2137	0.4045	0.4402	0.4381	0.2573	0.2050	0.2167
	ภูมิใจไทย	4	0.0627	0.0407	0.1160	0.0214	0.0299	0.0940	0.0641	0.0370	0.0969	0.0719	0.0614	0.1668	0.0595	0.0358	0.0965
	ประชาธิปัตย์	5	0.0117	0.0041	0.0081	0.0115	0.0012	0.0002	0.0129	0.0042	0.0082	0.0089	0.0060	0.0112	0.0128	0.0037	0.0070
	พลังประชารัฐ	6	0.0625	0.0180	0.0223	0.0179	0.0005	0.0004	0.0787	0.0195	0.0288	0.0357	0.0248	0.0168	0.0721	0.0164	0.0244
	ไทยสร้างไทย	7	0.0077	0.0109	0.0139	0.0086	0.0292	0.0514	0.0098	0.0098	0.0128	0.0022	0.0018	0.0012	0.0096	0.0130	0.0188
		Total	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

ตารางที่ 4 ข้อมูลอันดับของผลคะแนนการเลือกตั้งแบบบัญชีรายชื่อ (Dependent Variable) และข้อมูลจาก Twitter (Independent Variables) ทั้งที่เป็นข้อมูลดิบและร้อยละ

ผู้วิจัยได้ทำการนำเข้าข้อมูลอันดับของผลคะแนนการเลือกตั้งแบบบัญชีรายชื่อ (Dependent Variable) และร้อยละของข้อมูลจาก Twitter (Independent Variables) ตามตารางที่ 4 โดยทดลองนำเข้าตัวแปรอิสระหลายๆแบบ ทั้งแบบ 1 ตัวแปร, 2 ตัวแปร และ 3 ตัวแปร เพื่อหาสมการเชิงเส้นที่มีค่าสัมประสิทธิ์ (Coefficients) ที่ทำให้ Mean Squared Error (MSE) มีค่าน้อยที่สุด ผลที่ได้ตามตารางที่ 5 ซึ่งแสดงค่า Mean Squared Error (MSE) จากการนำเข้าตัวแปรอิสระๆหลายแบบ ทั้งหมด 21 กรณี

Mean Squared Error (MSE)

MSE	#Independent Variable	1					2	3
	Sentiment	-	Pos	Neu	Neg	Pos+Neu	Neg, Pos+Neu	Pos, Neu, Neg
	#Tweet	1.4834999	0.7800148	1.3033971	2.392686	1.2060886	0.991226016	0.733640916
	#Like	0.7898197	1.2905683	0.8623104	1.9466772	0.9284993	0.768494314	0.500350176
	#Retweet	0.78955	1.5932825	0.7701433	2.2920527	0.8908814	0.779889207	**0.256295584**

ตารางที่ 5 Mean Squared Error (MSE)

จากตารางที่ 5 พบว่าตัวแปรอิสระที่นำเข้าไปในสมการเชิงเส้นแล้วทำให้ค่า Mean Squared Error (MSE) มีค่าน้อยที่สุด (MSE = 0.2563) คือ การนำเข้า 3 ตัวแปรอิสระ ดังนี้ 1) จำนวนการ Retweet ข้อความที่เป็นเชิงบวก, 2) จำนวนการ Retweet ข้อความที่เป็นกลาง และ 3) จำนวนการ Retweet ข้อความที่เป็นเชิงลบ ซึ่งมีค่าสัมประสิทธิ์ (Coefficients) ดังสมการที่ 5

$$\text{Score Rank} = 6.1211 + 14.6953\,(\#\text{Retweet\_Pos}) - 27.7285\,(\#\text{Retweet\_Neu}) - 1.8143(\#\text{Retweet\_Neg}) \quad (5)$$

โดยเมื่อนำเข้าข้อมูลจำนวน Retweet แต่ละความรู้สึกผ่านสมการที่ 5 จะสามารถทำนายอันดับของผลการโหวตพรรคการเมือง ดังตารางที่ 6 ซึ่งค่อนข้างใกล้เคียงกับอันดับผลการเลือกตั้งที่เกิดขึ้นจริง

	Score Rank	#Retweet_ Pos	#Retweet_ Neu	#Retweet_ Neg	Predict
ก้าวไกล	1	0.5161	0.4437	0.1929	**1.05**
เพื่อไทย	2	0.1053	0.1959	0.1729	**1.92**
รวมไทยสร้างชาติ	3	0.2326	0.2137	0.4381	**2.82**
ภูมิใจไทย	4	0.0940	0.0969	0.1668	**4.51**
ประชาธิปัตย์	5	0.0002	0.0082	0.0112	**5.88**
พลังประชารัฐ	6	0.0004	0.0288	0.0168	**5.30**
ไทยสร้างไทย	7	0.0514	0.0128	0.0012	**6.52**
	Total	1.0000	1.0000	1.0000	

ตารางที่ 6 Predicted Value

ดังนั้นจึงสรุปได้ว่าข้อมูลจาก Twitter ที่สำคัญที่ใช้ในการอธิบายอันดับของผลการเลือกตั้งได้ใกล้เคียงที่สุด คือ จำนวนการ Retweet ทั้ง 3 ความรู้สึก (เชิงบวก, เป็นกลาง, เชิงลบ)

4. สรุปผลการวิจัย

4.1 อภิปรายผล

จากการเปรียบเทียบ Model ทั้ง 4 models พบว่า wangchanberta-base-att-spm-uncased model มีประสิทธิภาพสูงสุด เห็นได้จาก Accuracy, Precision, Recall, F1 score ที่มีค่าสูงสุดบนข้อมูล Test set ทั้ง 2 ชุด เนื่องจาก model นี้ถูก train บนข้อความภาษาไทยที่มีขนาดใหญ่ที่สุดในบรรดาทั้ง 4 โมเดล ซึ่งปริมาณข้อมูลที่ป้อนให้กับ Model ในการเรียนรู้เป็นปัจจัยสำคัญที่ทำให้ wangchanberta-base-att-spm-uncased model มีประสิทธิภาพสูงสุด และถูกคัดเลือกไปใช้ในการทำนายจำแนกความรู้สึกของข้อความ Twitter ในกรอบการวิเคราะห์ต่อไป

จากการวิเคราะห์หาข้อมูลเชิงลึก (Insight) ในระดับผู้ใช้ และในระดับข้อความ พบ insight ที่น่าสนใจ ดังนี้

<u>User Level</u>

- ผู้ใช้ทั้งหมด 22,979 คน (Usernames) ส่วนใหญ่เป็นกลุ่ม Lover 60% รองลงมาเป็นกลุ่ม Attacker 28% และ Both 12%
- แม้ว่าผลการเลือกตั้งที่เกิดขึ้นจริง คะแนนพรรคก้าวไกลจะมาเป็นอันดับหนึ่ง แต่ผู้ที่ใช้ Twitter ที่โพสต์ข้อความส่วนใหญ่นั้นชื่นชอบพรรคเพื่อไทยมากที่สุด รองลงมาเป็นพรรคก้าวไกลและพรรครวมไทยสร้างชาติ และพรรคที่มีผู้ใช้โพสต์ข้อความโจมตีมากที่สุด คือ พรรครวมไทยสร้างชาติ รองลงมาเป็นพรรคเพื่อไทย และพรรคภูมิใจไทย

<u>Tweet Level</u>

- หากพิจารณาแค่จำนวนข้อความ (No. of Tweet) จำนวน 102,997 Tweets พรรคเพื่อไทยเป็นพรรคที่ถูกกล่าวถึงมากที่สุด สะท้อนได้จากจำนวนผู้ใช้ (User) ที่มีจำนวนคนชื่นชอบพรรคเพื่อไทยมากที่สุด
- แต่เมื่อพิจารณาจำนวนการได้รับการกดถูกใจ (No. of Like) และจำนวนการ Retweet ที่มีมากถึง 3.46 ล้าน และ 4.21 ล้าน ตามลำดับ ซึ่งทั้ง No. of Like และ No. of Retweet ที่มากที่สุดตกเป็นของพรรคก้าวไกล ซึ่งสอดคล้องกับผลการเลือกตั้งที่เกิดขึ้นจริง จากสมมติฐานของผู้วิจัยที่เชื่อว่าผู้ใช้ Twitter มีมากกว่า 22,979 คนแน่ๆที่สนใจการเมือง แต่ไม่ได้แสดงออกผ่านการโพสต์ข้อความ แต่จะแสดงออกผ่านการกดถูกใจ (Like) หรือการแชร์ข้อความ (Retweet) ที่สนใจ เห็นได้จาก No. of Like และ No. of Retweet ที่มีมากกว่าจำนวนข้อความ (No. of Tweet) ถึง 30-40 เท่า ดังนั้นจึงเป็นไปได้ที่ No. of Like และ No. of Retweet จะสะท้อนผลการเลือกตั้งที่แท้จริงได้มากกว่า

<u>Multiple Linear Regression (MLR)</u>

- จากการทดลองประยุกต์ใช้ Multiple Linear Regression (MLR) เพื่อหาปัจจัยที่ใช้ในการอธิบายอันดับของผลการเลือกตั้ง ผลลัพธ์ที่ได้ คือ จำนวนการ Retweet ทั้ง 3 ความรู้สึก (เชิงบวก, เป็นกลาง, เชิงลบ) สามารถอธิบายอันดับของผลการเลือกตั้งได้ใกล้เคียงที่สุด ซึ่งยืนยันความเชื่อที่ว่า No. of Retweet จะสะท้อนผลการเลือกตั้งที่แท้จริงได้มากกว่า No. of Tweet และจะยิ่งทำนายผลได้ดียิ่งขึ้นเมื่อมีการเพิ่มข้อมูลเกี่ยวกับความคิดเห็น (เชิงบวก, เป็นกลาง, เชิงลบ) เข้าไปด้วย

4.2 ปัญหาและอุปสรรค

- ข้อมูลจาก Twitter ที่ใช้ในการวิเคราะห์ครั้งนี้ ตั้งแต่วันที่ 1 ม.ค. 2566 ถึง 10 เม.ย. 2566 เท่านั้น ยังขาดข้อมูลในช่วงวันที่ 11 เม.ย. 2566 ถึง 13 พ.ค. 2566 ซึ่งเป็นข้อมูลใกล้วันเลือกตั้งมากกว่า และข้อมูลยิ่งใกล้วันเลือกตั้งก็จะสะท้อนผลการเลือกตั้งที่แท้จริงได้มากกว่า เหตุที่งานวิจัยในครั้งนี้รวมรวบข้อมูลมาไม่ครบ เนื่องจากมีข้อจำกัดทางด้านการ scrape ข้อมูลจาก Twitter เพราะทางผู้บริหาร Twitter มีการเปลี่ยนแปลงนโยบายการแสดงผล ทำให้ Library ที่ใช้ไม่สามารถดึงข้อมูลในช่วงเวลาดังกล่าวได้

4.3 ประโยชน์

ผลสรุปการวิจัยนี้มีประโยชน์ต่อผู้ที่สนใจความคิดเห็นของสาธารณะชนทางด้านการเมือง ไม่ว่าจะเป็นผู้ที่มีหน้าที่บริหารพรรคการเมือง, ผู้กำหนดนโยบาย หรือผู้ที่สนใจทั่วไป สามารถนำข้อสรุปที่ได้ไปใช้ต่อยอดในการประเมินสถานการณ์การเลือกตั้งได้ เพื่อใช้ในการตัดสินใจทำเคมเปญหาเสียง หรือกำหนดทิศทางนโยบาย หรือแม้แต่การประยุกต์ใช้กรอบการวิเคราะห์ (Framework) ความคิดเห็นของผู้ใช้ Twitter ในแง่มุมอื่นๆที่ไม่ใช่แค่มุมทางการเมืองเพียงอย่างเดียว แต่สามารถนำกรอบการวิเคราะห์ในงานวิจัยนี้ไปสำรวจความคิดเห็นต่อเรื่องอื่นได้ๆ เช่น

- บริษัทอาจใช้กรอบการวิเคราะห์นี้เพื่อสำรวจความคิดเห็นของผู้ใช้เกี่ยวกับผลิตภัณฑ์หรือบริการของตน
- องค์กรไม่แสวงหาผลกำไรอาจใช้กรอบการวิเคราะห์นี้เพื่อสำรวจความคิดเห็นของผู้ใช้เกี่ยวกับประเด็นทางสังคม
- หน่วยงานรัฐบาลอาจใช้กรอบการวิเคราะห์นี้เพื่อสำรวจความคิดเห็นของผู้ใช้เกี่ยวกับนโยบายต่างๆ

โดยรวมแล้ว ผลสรุปการวิจัยนี้ให้กรอบการวิเคราะห์ที่มีประสิทธิภาพและยืดหยุ่นสำหรับสำรวจความคิดเห็นของผู้ใช้ Twitter

4.4 ข้อเสนอแนะ

- เพื่อวัดประสิทธิภาพของกรอบการวิเคราะห์ (Framework) และความยืดหยุ่นในการประยุกต์ใช้ ควรทำการทดสอบย้อนกลับ (Back Test) ในข้อมูลชุดอื่นๆ เช่น การทดสอบกับข้อมูล Twitter ในช่วงเวลาการเลือกตั้งผู้ว่ากรุงเทพมหานคร เป็นต้น

- ด้วยประสิทธิภาพ wangchanberta-base-att-spm-uncased model ที่มีค่า Accuracy บนข้อมูล Tweet Test Set อยู่ที่ 76% ซึ่งใกล้เคียงกับงานวิจัยอื่นๆที่ได้อ้างอิงมาก่อนหน้านี้ แต่การนำผลที่ได้ไปใช้ต่อใน Framework ที่เกี่ยวกับการ Retweet ซึ่งเป็นการทวีคูณผลลัพธ์ เช่น ทำนายผิดแค่ 1 Tweets จากเชิงลบ ทำนายผิดเป็นเชิงบวก ในแง่ของการนับจำนวน Tweet การทำนายผิดครั้งนี้จะนับแค่ 1 แต่ถ้าข้อความนี้ถูก retweet ไปเป็นหมื่นครั้ง ก็จะเป็นการทวีคูณความผิดพลาดได้ ดังนั้นจะต้องระมัดระวังในเรื่องของความผิดพลาดของ Model ที่ใช้ใน Framework

- อาจทดลองใช้ Logistic Regression Model แทนการใช้ Multiple Linear Regression (MLR) ในการหาความสัมพันธ์ระหว่างอันดับของผลการเลือกตั้งและข้อมูลจาก Twitter เนื่องจากงานวิจัยนี้ยังไม่ได้ทำการทดสอบสมติฐานความสัมพันธ์เชิงเส้นของตัวแปร ซึ่งข้อมูลชุดนี้อาจจะไม่ได้มีความสัมพันธ์เป็นเส้นตรงก็ได้

5. เอกสารอ้างอิง

[1] Schmidt, T., Fehle, J., Weissenbacher, M., Richter, J., Gottschalk, P., and Wolff, C., 2022, "Sentiment Analysis on Twitter for the Major German Parties during the 2021 German Federal Election," *In Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, Potsdam, Germany. pp. 74–87

[2] Sharma, P., & Moh, T.-S. (2016). Prediction of Indian election using sentiment analysis on Hindi Twitter. 2016 IEEE International Conference on Big Data (Big Data). IEEE.

[3] Macrohon, J. J. E., Villavicencio, C. N., Inbaraj, X. A., & Jeng, J.-H. (2022). A semi-supervised approach to sentiment analysis of tweets during the 2022 Philippine presidential election. Information (Basel), 13(10), 484. doi:10.3390/info13100484

[4] Ramadhan, WP, Astri Novianty, S.T., M.T & Casi Setianingsih, S.T.,M.T. (2017). Sentiment Analysis Using Multinomial Logistic Regression, Renewable Energy and Communications (ICCEREC).

[5] Lowphansirikul, L. et al., 2021, "WangchanBERTa: Pretraining transformer-based Thai language models," arXiv [cs.CL]. Available at: http://arxiv.org/abs/2101.09635.

[6] Suriyawongkul, A., Chuangsuwanich, E., Chormai, P., and Polpanumas., C., 2019. PyThaiNLP/wisesight-sentiment [Online], Available: https://github.com/PyThaiNLP/wisesight-sentiment. [20 April 2023]

[7] JustAnotherArchivist., snscrape: A social networking service scraper in Python [Online], Available: https://github.com/JustAnotherArchivist/snscrape. [5 April 2023]

[8] Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., & Chormai, P. (2016, Jun 27). PyThaiNLP: Thai Natural Language Processing in Python. Zenodo. http://doi.org/10.5281/zenodo.3519354

[9] Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. EMNLP.

[10] Limkonchotiwat, P., Phatthiyaphaibun, W., Sarwar, R., Chuangsuwanich, E., & Nutanong, S. (2020). Domain Adaptation of Thai Word Segmentation Models using Stacked Ensemble. EMNLP.

[11] สำนักงานคณะกรรมการเลือกตั้ง., รายงานผลการเลือกตั้ง ส.ส. ปี พ.ศ.2566 อย่างเป็นทางการ [Online], Available: https://official.ectreport.com/overview. [20 May 2023]