ANALYZING TWO OPPOSING OPINIONS FROM SOCIAL MEDIA THROUGH STATIC AND CONTEXTUALIZED WORD EMBEDDINGS

Miss Wassakorn Sarakul

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts in Linguistics
Department of Linguistics
FACULTY OF ARTS
Chulalongkorn University
Academic Year 2022

การวิเคราะห์ความเห็นสองขั้วจากสื่อสังคมผ่านการฝังคำแบบสถิตและ แบบอิงบริบท

น.ส.วรรษกร สาระกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Thesis Title	ANALYZING TWO OPPOSING OPINIONS FROM SOCIAL MEDIA THROUGH STATIC AND CONTEXTUALIZED WORD EMBEDDINGS
By	Miss Wassakorn Sarakul
Field of Study	Linguistics
Thesis Advisor	Associate Professor ATTAPOL THAMRONGRATTANARIT, Ph.D.

Accepted by the FACULTY OF ARTS, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Arts

.... Dean of the FACULTY OF ARTS
(Associate Professor SURADECH CHOTIUDOMPANT, Ph.D.)

THESIS COMMITTEE

.... Chairman
(Associate Professor Wirote Aroonmanakun, Ph.D.)

.... Thesis Advisor
(Associate Professor ATTAPOL THAMRONGRATTANARIT, Ph.D.)

.... External Examiner
(Can Udomcharoenchaikit, Ph.D.)

วรรษกร สาระกุล : การวิเคราะห์ความเห็นสองขั้วจากสื่อสังคมผ่านการฝังคำแบบสถิต และ แบบอิงบริบท. (ANALYZING TWO OPPOSING OPINIONS FROM SOCIAL MEDIA THROUGH STATIC AND CONTEXTUALIZED WORD EMBEDDINGS) อ.ที่ปรึกษาหลัก : รศ. ดร.อรรถพล ธำรงรัตนฤทธิ์

วิทยานิพนธ์เล่มนี้ศึกษาการวิเคราะห์ความเห็นสาธารณะจากข้อความในสื่อสังคมออนไลน์โดยใช้การฝังคำแบบสถิตและแบบอิงบริบท คำถามวิจัยของวิทยานิพนธ์นี้คือ การ pre-trained ส่งผลต่อการฝังตำแต่ละแบบอย่างไรเมื่อพิจารณาความแตกต่างของขนาดข้อมูลเปรียบเทียบกัน การฝังคำแบบใดที่มีประสิทธิภาพสูงสุดในการแยกความเห็นสองขั้วออกจากกัน และการฝังคำทั้งสองแบบมีพฤติกรรมต่างกันอย่างไรในการวิเคราะห์ความเห็นสาธารณะ ผลการศึกษาพบว่าการ pre-trained ส่งผลต่อคุณภาพของการฝังคำแบบสถิตเมื่อข้อมูลมีขนาดเล็ก แต่การ pre-trained กลับสร้างความแปรปรวนในข้อมูลขนาดใหญ่ ดังนั้นในข้อมูลขนาดใหญ่การใช้การฝังคำแบบสถิตโดยไม่ pre-trained จะให้ประสิทธิภาพสูงสุด การวิเคราะห์ความเห็นสาธารณะเหมาะกับข้อมูลขนาดใหญ่โดยใช้การฝังคำแบบสถิตโดยไม่ pre-trained มากกว่าการฝังคำแบบอิงบริบท เพราะการฝังคำแบบสถิตสามารถจับความสัมพันธ์ทางโครงสร้างประโยคซึ่งเป็นประโยชน์ต่อการวิเคราะห์ชนิดนี้ ส่วนการฝังคำแบบอิงบริบทจับความสัมพันธ์ทางความหมายได้ดีกว่า แต่ความสัมพันธ์ชนิดนี้ไม่เหมาะกับการใช้วิเคราะห์ความเห็น ดังนั้น แนะนำให้ใช้การฝังคำแบบสถิตทำนายฝั่งความเห็นแล้วใช้การฝังคำแบบสถิตหาคำใกล้เคียงกับคำที่ต้องการหาความเห็นมาใช้วิเคราะห์ความเห็นจะเป็นวิธีที่ดีที่สุด

สาขาวิชา ภาษาศาสตร์ ลายมือชื่อนิสิต ....
ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6380037122 : MAJOR LINGUISTICS
KEYWORD: Word Embedding, Word2vec, WangchanBERTa, Opinion Analysis

Wassakorn Sarakul : ANALYZING TWO OPPOSING OPINIONS FROM SOCIAL MEDIA THROUGH STATIC AND CONTEXTUALIZED WORD EMBEDDINGS. Advisor: Assoc. Prof. ATTAPOL THAMRONGRATTANARIT, Ph.D.

Public opinion analysis plays a vital role in various domains, such as marketing and politics. With the increasing volume of text data available through the internet and social media, efficient text-based analysis methods have become crucial. This study explores the application of static and contextualized word embeddings in word-based opinion analysis. The research questions focus on the impact of pre-training on static word embeddings, the efficacy of static and contextualized word embeddings in delineating opposing opinions, and the behavioral differences between the two embedding types. The findings suggest that pre-training improves embedding quality in small-sized datasets but may introduce noise in large-sized datasets. Additionally, word-based opinion analysis is more suitable for large-sized datasets, with un-pre-trained static word embeddings demonstrating superior performance. Static word embeddings are preferred over contextualized word embeddings due to their ability to capture syntactical relationships, while contextualized word embeddings provide semantic-related similar words. To apply both embedding types effectively, the study recommends using contextualized sequence embedding to predict the corpus, training word2vec on the predicted corpus, and analyzing the corpus based on the most similar words from the word2vec model.

Field of Study: Linguistics Student's Signature ....
Academic Year: 2022 Advisor's Signature ....

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt appreciation to my advisor, Associate Professor Dr. Attapol Thamrongrattanarit, for his unwavering support of my Master's studies and research, as well as his patience, encouragement, excitement, and extensive expertise. His advice was invaluable during the research and writing of this thesis.

In addition to my advisor, I'd like to thank the other members of my thesis committee: Associate Professor Wirote Aroonmanakun and Dr Can Udomcharoenchaikit, for their support and insightful suggestions.

My sincere thanks also go to Lalita Lowphansirikul, who kindly assists Wangchanberta's technological challenges.

I thank my fellow labmates in the NLP laboratory: Apiwat Jaroonpol, Narongkorn Panitsrisit, Pakawadee Pamorn, Jin Cheevaprawatdomrong, Pawitsapak Akarajaradwong, for their comments and support.

Last but not least, I would like to thank my parents, Sampop Sarakul and Vipa Sarakul, for their invaluable support in every aspect.

Wassakorn Sarakul

TABLE OF CONTENTS

Page

List of tables

List of figures

INTRODUCTION

Public opinion is a highly relevant and widely studied subject, as it provides numerous advantages. For example, companies constantly monitor the trends in the market to optimize their sales strategies, and politicians seek to understand the concerns of their constituents to craft effective campaign messages. The examination of public opinion is often facilitated through the analysis of text, which serves as a rich source of information about the prevailing sentiments and issues in society. As such, text-based analysis has become an increasingly important tool for understanding public opinion and its various applications.

The rapid growth of the internet and social media has led to a proliferation of texts being produced and published on a daily basis. In 2020, it was estimated that 500,000 tweets were posted every minute (Vish, 2020). These texts provide a valuable resource for exploring topics such as bias, stereotypes, sentiment, and opinion. Policymakers, for instance, may seek to track public reactions to their policies in real time, while marketers may want to gauge the impact of their marketing campaigns as quickly as possible. This information can then be used to adjust policies and campaigns promptly, leading to more effective outcomes. Unlike traditional methods such as surveys, social media text offers the advantage of being created by users in real-time, with a large volume and greater emotional expressiveness. As a result, it is increasingly being recognized as an excellent source for monitoring policy developments and public opinion ("How suitable is Twitter for monitoring policy developments?," 2021)

However, relying solely on human labor for analyzing these vast amounts of text is not feasible for several reasons. Firstly, the sheer volume of text presents a significant challenge regarding the amount of human labor required for analysis. Secondly, the fast-paced nature of text generation can result in decreased quality of analysis due to the limited amount of time available for analysis. To overcome these challenges, it is necessary to employ algorithms that can efficiently process and gain insights from large amounts of data. Using algorithms ensures that the analysis remains thorough, accurate, and up-to-date, even as the text volume continues to grow.

An algorithm must deeply understand the language in question to analyze language effectively. In natural language processing (NLP) systems, the smallest unit of analysis is typically a word. To represent the meaning of words, algorithms create a concept known as word representation, word vector, or word embedding. This is essentially a vector that encodes the meaning of a word within itself. The underlying idea behind word embedding is that the meaning of a word is derived from its context, meaning that the way a word is used expresses its sense (Harris, 1954). As a result, similar words will have similar word vectors because they tend to be used in similar contexts. This understanding of word meaning is crucial for accurately analyzing and interpreting language and forms the foundation for many NLP algorithms.

The concept that similar words have similar word vectors is the cornerstone of opinion mining, a widely used method in NLP. This task can be called word-based opinion analysis. This method typically begins with training word embeddings on a corpus of text, embedding the fact of how words appear in the corpus. The analysis is then conducted by comparing word vectors, for example, to determine if two groups have different opinions on a topic such as immigration. This process is done by splitting the text into two groups based on the individuals who generated it, then training word embeddings on each split text. If two groups have different opinions, their usage of a word like “immigrant” will likely differ, which can be observed by comparing the word vectors for “immigrant”. This method has been successfully applied in several studies, such as examining gender bias (Garg et al., 2018; Nelson, 2021), racism (Tahmasbi et al., 2021), and polarized political opinions (McCarthy et al., 2021).

In the field of NLP, there are two main types of word representation models: static word embeddings and contextualized word embeddings. Static word embeddings represent a single lexicon with a single vector, and this type of representation has the limitation of being uncontextualized. Regardless of the context in which a word appears, its representation remains unchanged. However, it is widely recognized that word representation should be context-sensitive, as the meaning of a word can change depending on the surrounding terms. As a result, contextualized word embeddings have been developed to better capture the meaning of words in context.

The static word embedding represents a single lexicon with a single vector, which cannot adjust a word's representation based on its context. A popular language model for training word embeddings is word2vec, a feedforward neural network that learns word embeddings by predicting the surrounding context or using the context to predict the focus word (Mikolov et al., 2013). In contrast, contextualized word embedding creates a separate vector representation for each token, allowing the same lexicon to have different representations in different contexts. The state-of-the-art contextualized word embedding model is Bidirectional Encoder Representations Transformer (BERT), which uses a self-attention mechanism to generate word representations (Devlin et al., 2018). The model allows for considering the relationships between the focus token and other tokens in the sequence, providing a complete understanding of how the word is used in context. This feature should allow for clustering words by their senses or analyzing polysemous words.

Despite the significant advancements in contextualized word embeddings, there remains a scarcity of research on their application in word-based opinion analysis. To date, no study has been conducted to compare the efficacy of the contextualized word embedding models in opposing opinion analysis. This gap in the literature highlights the need for further investigation into the use of contextualized word embeddings in the analysis of opposing opinions.

The objectives of this study are:

- To analyze the effectiveness and behavior of the output of static and contextualized word embeddings when applied to varying amounts of data.
- To gain best practices for when and how to apply static and contextualized word embeddings in the analysis of two opposing opinions.

To achieve this goal, I use two well-known algorithms for word embedding generation, word2vec (Mikolov et al., 2013) and WangchanBERTa (Lowphansirikul et al., 2021). The former generates static word embeddings, while the latter generates contextualized ones. To perform a word-based opposing opinion analysis, I create two corpora of Twitter text: one concerning the 2020-2021 Thai democratic protest and the other from the opposition.

The research questions are:

- How does the pre-training process affect the output of the static word embeddings in the word-based opinion analysis task?
- Which of the two embeddings is more useful in delineating the opposing opinion in two opposing focus corpora?
- How do static and contextualized word embeddings behave differently in word-based analysis?
- To gain best practices for when and how to apply static and contextualized word embeddings in the analysis of two opposing opinions.

The major arguments, theories or hypotheses are:

- Static word embeddings will overfit more to the data and yield nearest words that reflect related concepts expressed in the text, but contextualized word embeddings yield nearest words that reflect lexical relations or syntactic relations to the focus words because they have been pretrained on a massive raw dataset.
- When given enough data, static word embeddings might be favorable because there are enough repeated occurrences of words. In contrast, when given a small amount of data, contextualized word embeddings might be favorable because the model already learns part of the language through pretraining.

LITERATURE REVIEW

Before proceeding to the experiment, the related works of literature will be presented in this section. This chapter will begin with word embeddings, which are fundamental for word-based analysis. The differences between static word embedding and contextualized word embedding are illustrated here. Then, the studies that utilized word embedding in social sciences will be presented.

Word Embedding

In natural language processing, having the representation of meaning is crucial. Word representation or word embedding encodes syntactic and semantic meaning into a vector. For example, from the perspective of syntactic meaning, the word “car” is a noun. It can be an object of a verb, e.g., James usually drives his black car on Monday. Besides the structural aspect of the word, “car” also has a semantic meaning. Several ways can explain the semantic meaning. One of them is the lexical relation or how the word is related to other words. For example, the hypernym of “car” is “vehicle”, “car” is the hyponym of “vehicle”, and “motorcycle” is the cohyponym of “car”.

Word embedding is created based on the distributional hypothesis. The meaning of a word is derived from its surrounding words. Alternatively, how the word is used with other words shows the sense (Harris, 1954). Given the following examples of the word “pad thai”:

a. Many people always order pad thai when visiting Thai restaurants.
b. Pad thai is a favorite dish for many people.
c. Pad Thai is typically made with Rice Noodles.

The meaning of “pad thai” is derived from its context (Jurafsky & Martin, 2021). From the examples, if I do not know what is “pad thai”, I will try to get its meaning from the fact that “pad thai” can be an object of “order” (a.). In addition, it associates with “restaurants” (a.). Furthermore, it can be a meal because it occurs with a “dish” (c.). Lastly, one of pad thai’s ingredients is noodles. Without a clear definition of “pad thai”, I have an idea of its meaning from its co-occurrence.

Suppose from the context words above, and they also have their context as follows:

d. James ordered fried rice at Tim’s restaurant yesterday.

e. Steak is the main dish for today's dinner.

f. Ramen is a Japanese noodle soup.

From all the example sentences, I have an idea that “pad thai” is a kind of meal similar to fried rice, steak, and ramen since they have a similar environment. This is the idea of how word meaning derived from its context.

The word embedding is a vector representing a word. The Word vector’s dimension varies from 50 to 1,000 dimensions. Each dimension contains a real number that is also called a weight.

There are several algorithms to create word embedding; for instance, word2vec (Mikolov et al., 2013), fasttext (Bojanowski et al., 2017), GloVe (Pennington et al., 2014). Word embedding can be divided into static word embedding and contextualized word embedding.

Word embedding similarity

When converting the meaning into a vector, we can compare the meaning by comparing the word vectors or word embeddings. The similarity between word embedding is calculated from the cosine similarity. The cosine angle between the vectors shows how similar these vectors are. The formula for cosine similarity is shown in Equation 1. (Kaveh-Yazdy & Zarifzadeh, 2021)

$$Sim\left(V_x, V_y\right) = Cosine(\theta) = \frac{V_x \cdot V_y}{\|V_x\|\|V_y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt[2]{\sum_{i=1}^{n} x_i}\,\sqrt[2]{\sum_{i=1}^{n} y_i}}$$

Equation 1

The distance between word embedding is the flip side of the similarity. It can be calculated in Equation 2.

$$Distance\left(V_x, V_y\right) = 1 - \ Sim\left(V_x, V_y\right)$$

Equation 2

The most similar word is the word that has the highest cosine similarity with the current word $\mathrm{V_x}$.

Static Word Embedding

Static word embedding uses one vector to represent one word regardless of its sentence positions. Considering the following three examples,

g. “The poor bird couldn't fly because it had a broken wing.” (Dictionary, 2022b)

h. The gunman shot the princess and ran to the east wing of the palace.

i. “The extreme right-wing of the party has dominated the discussion.” (Dictionary, 2022b)

The word “wing” presented above have a different meaning. From sentence g., “wing” is the part of a bird’s body that helps the bird fly. On the other hand, “wing” is a part of the building in sentence h. In contrast, it is the political side in sentence i. Nevertheless, the vector representations of these three wings are the same, so this word embedding is context-independent. The same vector representation regardless of the environment.

The most notable algorithm to train static word embedding is word2vec by Mikolov, Chen, et al. (2013). In summary, word2vec has two alternatives; Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model (Skip-gram). (Figure 1) The first model creates word embedding by using the context words’ embedding to predict the current word, while the latter uses the current word’s embedding to predict the context. According to Mikolov, Chen et al. (2013), the continuous Skip-gram model is more effective than CBOW, and Skip-gram is more widely used. Thus, I will briefly explain CBOW, then continue to demonstrate the Skip-gram in detail.

Given the piece of text as the training data:

“Domestic cats are valued by humans for companionship …” ("Cat," 2023)

$$w_{t-2} \quad w_{t-1}\ w_t \quad w_{t+1} w_{t+2}$$

Word2vec trains word embedding from the running text. Given the width of the context window is 2, w_t is the target word or the focus word, and $\mathbf{w_{t-2}}$, $\mathbf{w_{t-1}}$, $\mathbf{w_{t+1}}$, w_{t+2} is the context words.

Figure 1 shows the two architectures. In the Continuous Bag-of-Words Model (CBOW), the vector of context words is summed to predict w_t. Alternatively speaking, the context words are the inputs w_I And the target word is an output w_O in CBOW. On the contrary, the target word is an input w_I to predict the contexts w_O in Skip-gram.

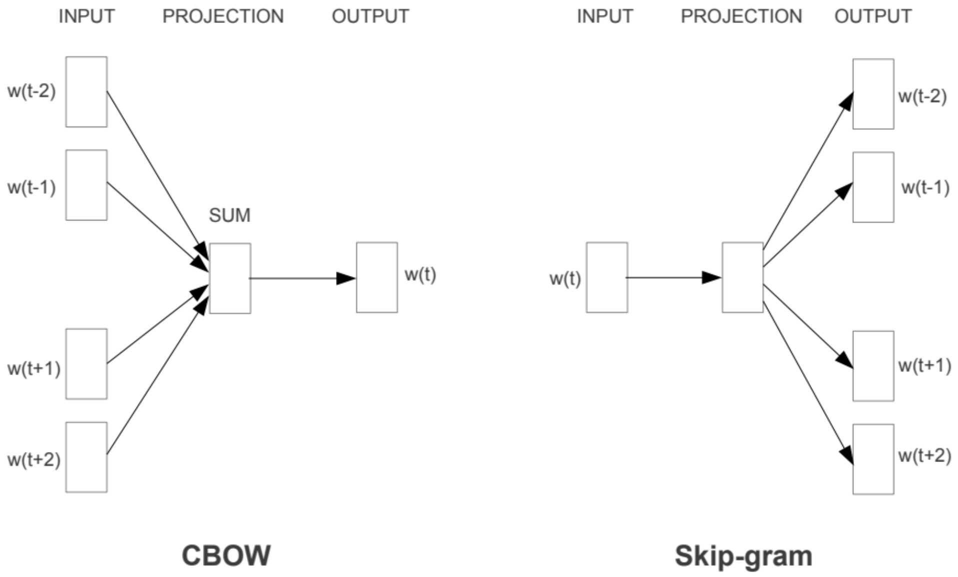


Figure 1: Two Model Architeture of Word2Vec (Mikolov et al., 2013)

The intuition behind the word2vec is that if w_I and w_O appear in the text together, the probability that w_O occurs given w_I should be high.

$$p(w_O|w_I) \; = \; \frac{exp(v'_{w_O} \top v_{w_1})}{\sum_{w=1}^{W} exp\,(v'_w \top v_{w_I})}$$

Equation 3

$$p(w_O|w_I) \; = \; log\, \sigma(v'_{w_O} \top \, v_{w_I})$$

Equation 4

Equation 3 defines the probability that w_O occurs given w_I. v_w is the vector representation of word w. W is the total number of words in the vocabulary. Equation 3 can be interpreted as the similarity between the vector representation of w_I and w_O. The words that appear together should have a more similar vector representation; thus, the fraction of Equation 3 will be high. Therefore, the objective of Skip-gram is to find the vector representation that maximizes Equation 3. The $p(w_O|w_I)$ is also equal to Equation 4, where σ is the softmax function.

In addition, Skip-gram also has negative sampling. The idea of negative sampling is that the efficient representation can give a high probability to the correct context words and a low probability to incorrect context. The k incorrect contexts are drawn from the noise distribution $P_n(w)$. Thus, the objective function of Skip-gram is Equation 5, in which the first term is the probability of the correct context and the second term is the probability of incorrect context.

$$log\ \sigma(v'_{w_O} \top v_{w_I}) + \sum_{i=1}^{k} E_{w_i \sim P_n(w)}\left[log\ \sigma(-v'_{w_O} \top v_{w_I})\right]$$

Equation 5

Lastly, Skip-gram employs subsampling of frequent words, which is discarding the frequent word with the probability $P(w_i)$ in Equation 6. Considering the training example, frequent words such as “by” and “for” occur substantially frequently. However, they have less information than less frequent words such as “companionship”. Thus, to learn the representation of “humans”, it is more valuable to learn the co-occurrence of less frequent words.

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

Equation 6

Skip-gram trains word embedding using stochastic gradient descent and backpropagation to get the word representation that maximizes the objective function described in Equation 5.

Word2vec has a linear property. The arithmetic operation can be done on word2vec. For example, the most similar vector to vector(“Berlin”) - vector(“Germany”) + vector(“France”) is vector(“Paris”). This is an example of semantic meaning encoded in a word embedding. The syntactic meaning is also encoded; for instance, vector(“quickly”) - vector(“quick”) + vector(“slow”) is close to vector(“slowly”). (Mikolov et al., 2013)

<u>Contextualized Word Embedding</u>

While static word embedding uses one vector to represent one word in the vocabulary, contextualized word embedding uses one vector to represent one token in the text based on the context. Here are the examples of the word “man”:

j. There is a man in a green shirt standing under the apple tree.

k. “Man is still far more intelligent than the smartest robot.” (Dictionary, 2022a)

l. “The phones are manned 24 hours a day.” (Dictionary, 2022a)

The term “man” appears three times in the given text. These tokens will have the exact vector representation for the static word embedding. On the other hand, for contextualized one, the representation of “man” differs because they occur in a different context. In sentence j., the “man” is a noun meaning “an adult male human being” (Dictionary, 2022a). The “man” in sentence k. is also a noun, but it means “the human race” (Dictionary, 2022a). Lastly, the “man” in l. is a verb meaning “to man something such as a machine or vehicle is to be present in order to operate it” (Dictionary, 2022a). As opposed to static word embedding with the exact representation of “man” regardless of the context, contextualized word embedding will give a new version of the word vector when the surrounding word changes.

The static word embedding has the disadvantage of the fixed representation and the limited length of context. As demonstrated in the Word2Vec model architecture, the number of contexts must be specified. Thus, it cannot consider the longer context and fail to capture the long-range dependency. It does not capture the whole meaning of the word.

Self-attention

There is an invention proposed to overcome these limitations. It is a self-attention mechanism proposed by Vaswani et al. (2017). Their model architecture consists of two parts; encoder and decoder. As presented in Figure 2, the left part is the encoder, and the right part is the decoder. The green box in Figure 2 is the transformer. The transformer can be both an encoder and a decoder. It is the self-attention (the orange box) stacked by the dense layer (the light blue box). The green box in Figure 2 presents one layer of the transformer.

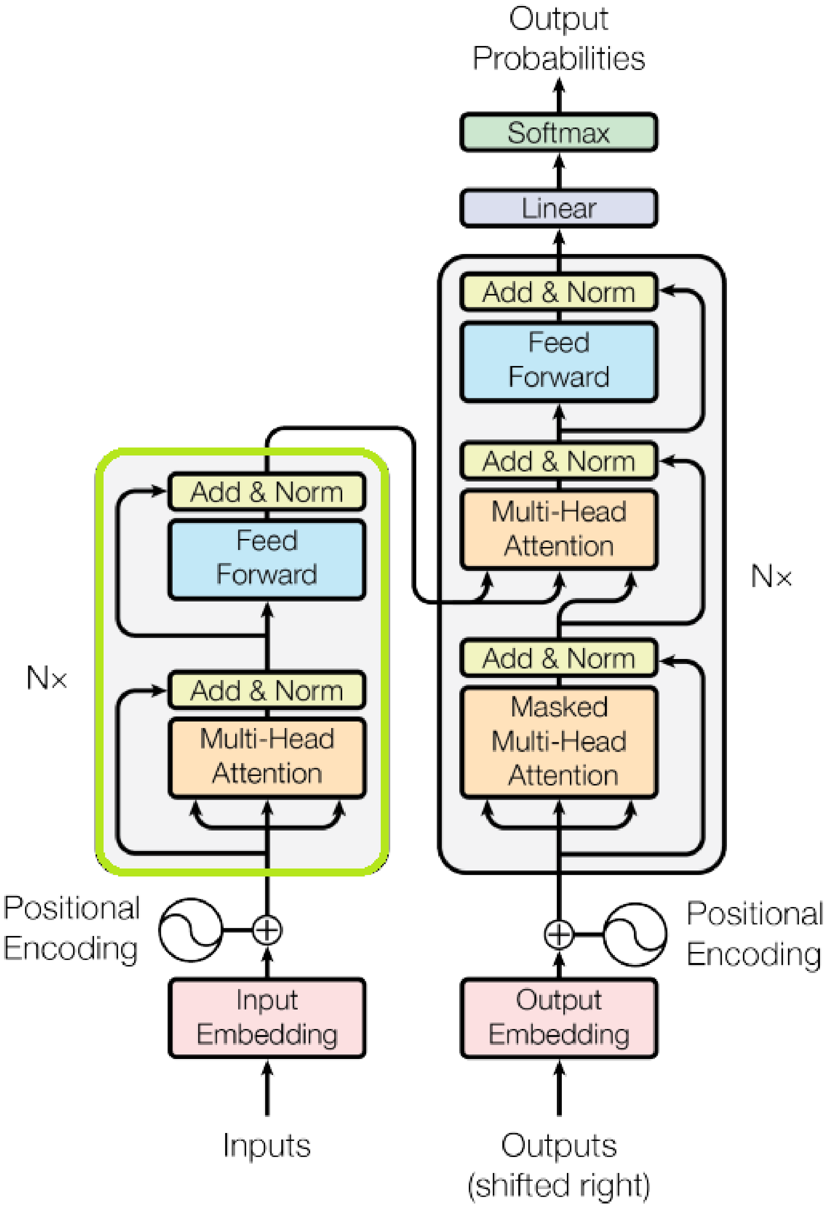


Figure 2: The transformer - model architecture (Vaswani et al., 2017)

The critical feature in the encoder is Multi-Head Attention. It learns more than one relation between the words in the sequence. The intuition behind it is the self-attention mechanism. To create the representation of a particular token, the token also relates to other tokens. The attention shows how important the other tokens in the sequence relate to the focus token. Figure 3 shows the relation between “making” and other tokens in the sequence. The edge shows the relation, and the value of the edge indicates the strength of the relation. In Figure 3, the word “making” has a strong relation to “more” and “difficult” aligning with the pattern “making ... more difficult”

(Vaswani et al., 2017). Word2Vec cannot capture this long-distance dependency because it may have a smaller context than the range, where “making” and “more difficult” co-occur. Utilizing attention makes representation encodes richer meaning.

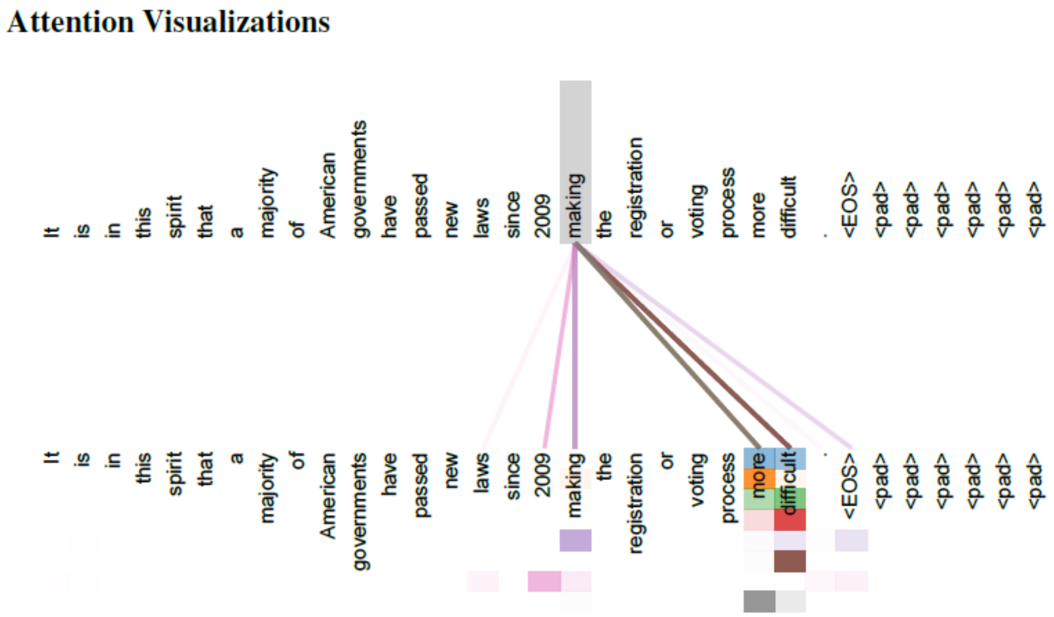

Attention Visualizations
It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>
It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>

Figure 3: Attention Example (Vaswani et al., 2017)

The attention is calculated from Scaled Dot-Product Attention. Firstly, the input sequence length n has the input embeddings $(x_1, x_2, \dots, x_n)$. These input embeddings are transformed into query q, key k, and value v. The calculation does not process token by token but uses matric operation. The input embeddings are stacked to form X matrix. Then, the learned weight matrix W^Q, W^K, and W^V projects X to matrix $Q, K,$ and V as shown in Equation 7, Equation 8, and Equation 9.

$$Q = XW^Q$$

Equation 7

$$K = XW^K$$

Equation 8

$$V = XW^V$$

Equation 9

Where $W^Q \in R^{d_{model} \times d_k}, W^K \in R^{d_{model} \times d_k}$, and $W^V \in R^{d_{model} \times d_v}$. (Vaswani et al., 2017)

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Equation 10

Where $Attention(Q,K,V) \in R^{N\times d_v}$. The interpretation of QK^T term is the comparison of each token to all other tokens in the input. Then, the term is scaled down by $\frac{1}{\sqrt{d_k}}$. The scaled dot-product is fed into the softmax function to create the weight. Each row of matrix $Attention(Q,K,V)$ correspond to each row of matrix X. Alternatively speaking, the representation of x_i is the weighted sum of all tokens in the sequence.

Because it is possible to have more than one relation among the tokens, Vaswani et al. (2017) employ multiple duplicates of attention called Multi-head attention networks. Each duplicate or head has its own parameter matrix W^Q,W^K, and W^V. Therefore,

$$head_i = Attention(Q_i, K_i, V_i)$$

Equation 11

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \dots, head_n)W^O$$

Equation 12

Where $W^O \in R^{hd_v\times d_{model}}$, and h is the number of heads. Finally, all heads are concatenated. (Figure 2) Then, it is connected with the residual connection and layer normalization.

$$FFN(x) = max(xW_1 + b_1)\,W_2 + b_2$$

Equation 13

Where x is the output from the previous sub-layer, again, the output from Equation 13 is connected to residual connection and layer normalization.

All the steps explained above form one layer of the transformer.

<u>*BERT Architecture*</u>

The state-of-the-art model architecture, BERT or Bidirectional Encoder Representation from Transformers, employs a self-attention mechanism. BERT uses 12 transformer layers in $BERT_{BASE}$ and 24 layers in $BERT_{LARGE}$. (Devlin et al.,

2018) The output from one transformer layer is passed on to the next transformer layer as an input.

BERT constrains the input sequence length not to exceed 512 tokens. The longer the input text is, the higher computational cost is required. Thus, BERT needs to constrain the input length to a sufficient range to acquire the context and not too long.

<u>*BERT Training*</u>

The power of BERT comes from two reasons; the first is the self-attention mechanism, and the second is the transfer learning from a substantial amount of data. Firstly, BERT acquires knowledge of the language in the pre-training process. Then, it is fine-tuned to adapt to the downstream task. I will demonstrate the pre-training first.

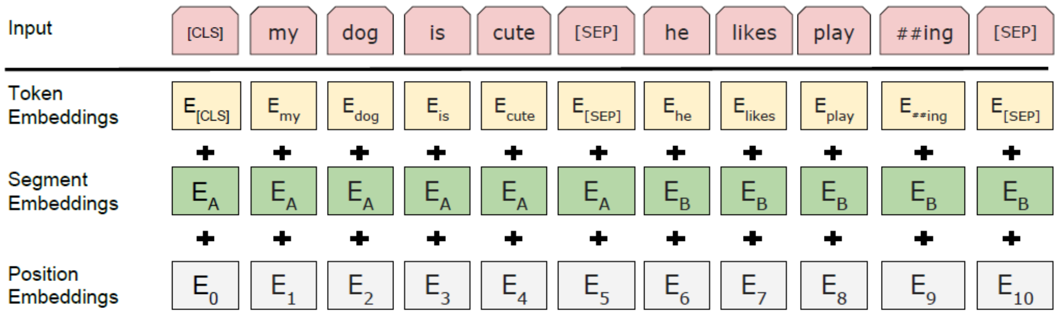


Figure 4: BERT input (Devlin et al., 2018)

BERT used a WordPiece tokenizer (Wu et al., 2016). This tokenizer tokenized the text into subwords. The example is in Figure 4. The word “playing” is tokenized into “play” and “# #ing”. The intuition behind the subword tokenization instead of word-level tokenization is that it requires a substantial amount of memory and time to cover as many word forms as possible. Here is a great example, the word “annoy” has several word forms, such as “annoying”, “annoys”, and “annoyed”. Likewise, other words also have several word forms. Suppose one word has 3 word forms, and there are 3 words (citation form) in the dictionary. As a result, there are 9 words in the dictionary if one word form is assigned to one entry. This way, the vocabulary size will grow extensively since one entry for every word form is needed. Splitting “annoying” into “annoy” and “# #ing” will reduce the size of vocabulary and save computational time and memory. According to the example of 3 words (citation form) in the dictionary, the vocabulary size is 6 (3 + 3). (HuggingFace, 2022)

The other reason is that the subword has a higher frequency than the “annoying” so BERT will get more samples for “annoy”, resulting in better representation. In addition, “annoying” has the composite meaning of “annoy” and “-ing”. Therefore, BERT should learn the representation of each subword. Moreover, subword tokenization allows BERT to handle the unknown word or the word that does not appear in the training sample. For instance, “transforming” is tokenized into “transform” and “# #ing”. “transforming” may not be in the training data, but it is more likely that “transform” is in the training data, and “# #ing” too. This tokenization method reduces the out-of-vocabulary problem when implementing the model in the downstream task. (HuggingFace, 2022)

WordPiece tokenizer (Wu et al., 2016) is the subword tokenizer. Schuster and Nakajima (2012) firstly introduce this tokenizer. The process begins with creating the inventory of all characters. Then, the model builds the language model for the training data from the inventory. They increment the word unit to the inventory by adding the pair of units in the inventory one by one. The criteria for choosing the pair of units is that the pair should increase the likelihood of the training data. The process will repeat until the number of word units reaches the predefined number or the increase in likelihood is below the predefined level. WordPiece uses this word inventory to tokenize the text by matching the word unit and the input text, and tokenizing them accordingly.

In addition to token embeddings, BERT also trains the segment embeddings to differentiate the sentence in the input sequence. As presented in Figure 4, there are two sentences. The [SEP] token is added at the end of each sentence. The segment embedding (the green box) E_A shows that the tokens preceding the [SEP] token and the [SEP] token belong to the first segment, and E_B shows that the tokens behind the previous [SEP] belong to the second segment. The [SEP] token encodes that the input text is a pair of segments in the segment embedding of the [SEP] token.

Lastly, the position embeddings are trained to use the token's position in sentences. The position embedding encodes the relative positional information of each token. The positional information is critical for languages because language is not the random order of the words, but words adjacent in particular order create the whole meaning of the sentence. According to the architecture of the transformer, the mechanism encodes the sequence in parallel, and no part deals with this information. The positional embedding acts as a signal to the model about the position of the input token.

The positional embedding is calculated from the sine and cosine functions. Each dimension in the embedding has a function to calculate the positional weight from the position. In other words, the positional weight is the function of the position of that token in the sequence. Therefore, a position in a sequence such as one can be transformed into a vector containing real numbers. The formulas are shown as follows:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{2i}/_{d_{model}}}\right)$$

Equation 14

$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{2i}/_{d_{model}}}\right)$$

Equation 15

Where pos is the position of the token, and i is the dimension of the positional encoding.

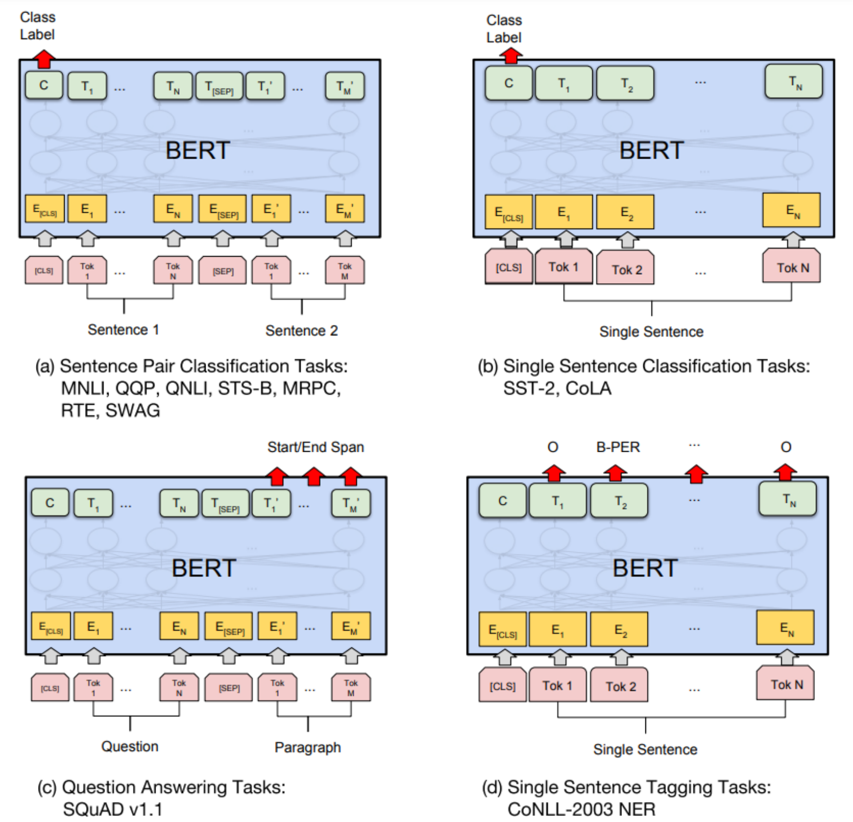
In conclusion, the input layer of BERT is the sum of token embeddings, segment embeddings, and position embeddings.

BERT training consists of two steps: pre-training and fine-tuning. The pre-training is to train the model with a tremendous amount of unlabeled text, while fine-tuning is to finetune the previously trained parameter with the downstream task such as sentiment prediction.

BERT has two pre-training tasks. The first task is the masked language model (MLM). This task is like the cloze test. Some words are removed, so participants need to fill in the blank with the correct answer. Intuitively, the participant needs to know the language to understand the context to find the correct answer. They need to know the vocabulary and its usage. The ability to do this task correctly shows that the participant understands the language. Therefore, BERT acquires the language by learning this task. To train the MLM task, fifteen percent of the tokens are randomly selected. Of the chosen tokens, 80 percent are replaced by [MASK], random incorrect tokens replace 10 percent, and the rest 10 percent are still unchanged. Only the final hidden layers $h_i$of these selected tokens are used to predict the original tokens. The cross-entropy losses are averaged within the sequence, and the gradient is calculated for the case of stochastic gradient descent. For batch gradient descent, the cross-entropy losses are averaged within the batch.

The second task is next sentence prediction (NSP). The objective of the training NSP task is to allow the [SEP] token to encode the fact that the input is a pair of sentences. In addition, many downstream tasks such as question answering, paraphrasing detection, and natural language inference need an understanding of the relation between sentences. Fifty percent of the input sequences are selected, and the next sentences are replaced by random sentences. The rest of the input sequences are left unchanged. Thus, the model will predict whether the following sentence is the real adjacent sentence or not. Therefore, the loss function of BERT is the sum of loss from MLM and loss from NSP.

BERT trains on the massive corpus. It achieves the state-of-art GLUE score, MultiNLI, SQuAD v1.1, and SQuAD v2.0 test. (Devlin et al., 2018)



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Figure 5: BERT Fine-tuning (Devlin et al., 2018)

BERT is compatible with several downstream tasks. It is easy to fine-tune by plugging specific input and output of that task into the model. Then, all the parameters are fine-tuned. Figure 5 shows the several downstream tasks and how input and output are different.

In the single sentence classification task (b) and single sentence tagging task (d), the input sequence does not have the [SEP] token. On the other hand, in the sentence pair classification task (a), e.g., natural language inference and question answering task (c), the input sequence has the [SEP] token because the input is a pair of sentences.

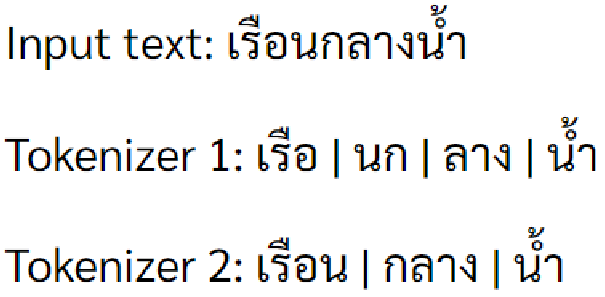
From the output side, the token representation of the [CLS] token is connected to the output layer for the sentence pair classification task (a) and the single sentence classification task (b). The [CLS] token acts as the sentence embedding in BERT. It encodes the entire sequence. While only [CLS] representation is used in (a) and b, more than one token representation is used in (c) and (d). In the question answering task (c), the [SEP] token indicates the end of the first sentence. Then, every token representation after the [SEP] token is fed into the classifier to predict whether the particular token is the start/end of the text span. In single sentence tagging task (d), the [CLS] shows that the model should begin classifying the next token. The token representation of each token is fed into the classifier to predict the labels. For example, it gives the label O to the token that is not in the part of the name entity and gives B-PER to the token that is the beginning of the PERSON name entity.

WangchanBERTa

Although BERT's descendant model has become a state-of-art in English, they still underperform (relative to English) in many languages especially low resource languages, due to several limitations.

The first limitation is the smaller amount of train data. BERT multilingual base model (mBERT) trains on the 104 languages Wikipedia pages. (Devlin et al., 2018) Compared to the training data for the English language of the BERT, the training data for each language is still substantially lower. The second limitation is that mBERT does not consider language-specific features. For example, some languages have different morphology from English.

The Thai language differs from English in some aspects. Firstly, Thai words are not separated by spaces—all words are adjacent to each other in sentences. Therefore, space cannot be used for word tokenization as in English. There is more than one possible result of tokenization, as shown in Figure 6. The exact input text results in different token lists with diverse meanings. The second point is that space acts as a sentence boundary in the Thai language, like the period in English. Thus, the model that does not consider this feature will underperform or yield the worst result.



Input text: เรือนกลางน้ำ

Tokenizer 1: เรือ | นก | ลาง | น้ำ

Tokenizer 2: เรือน | กลาง | น้ำ

Figure 6: Thai tokenization

WangchanBERTa is the transformer-based model for Thai language. Its advantages come from the diverse training dataset such as social medial posts from various platforms, news, articles, Thai Wikipedia dump, subtitles, book corpus, and translations from various corpora. The cleaned training dataset is 78 GB in total size, much larger than the Thai Wikipedia dump used in mBERT (515 MB). The variety of the data sources makes WangchanBERTa cover language in diversified topics. (VISTEC, 2021) In addition, WangchanBERTa’s dataset has formal language, informal language, spoken language, and written language, while that of mBERT is formal language and written language. Therefore, WangchanBERTa learns a complete picture of the Thai language more than mBERT.

WangchanBERTa implements preprocessing the text before training since the cleaned training data give better performance than the uncleaned dataset. The preprocessing scheme includes cleaning the text by removing HTML tags, empty parenthesis, brackets and replacing space with <_>. WangchanBERTa used the SentencePiece subword tokenizer (Kudo & Richardson, 2018), combining spaces with other tokens. Because space is a sentence or phrase boundary in Thai and acts as the full stop in English, a special symbol for space is needed to preserve the sentence boundary. There are 381,034,638 unique sentences in total after preprocessing as a result.

WangchanBERTa implements preprocessing the text before training since the cleaned training data give better performance than the uncleaned dataset. The preprocessing scheme includes cleaning the text by removing HTML tags, empty parenthesis, brackets and replacing space with <_>. WangchanBERTa used the SentencePiece subword tokenizer (Kudo & Richardson, 2018), combining spaces with other tokens. Because space is a sentence or phrase boundary in Thai and acts as the full stop in English, a special symbol for space is needed to preserve the sentence boundary. There are 381,034,638 unique sentences in total after preprocessing as a result.

The training process of WangchanBERTa begins with training the tokenizer. SentencePiece subword tokenizer trains on randomly selected 20,961,306 sentences. WangchanBERTa model architecture has 12 transformer layers, 768 hidden dimensions, and 12 attention heads. (Lowphansirikul et al., 2021) Lowphansirikul et al. (2021) adopt RoBERTa-base architecture (Liu et al., 2019). RoBERTa architecture is similar to BERT except removing the next sentence prediction, changing the masked token when training on different epochs, bigger batch size, and longer sequences. Due to the limitation of computing power, the max sequence length of WangchanBERTa is 419 tokens, whereas that of RoBERTa is 512 tokens. (VISTEC, 2021)

WangchanBERTa's result surpasses the strong baselines on the sequence and token classification. WangchanBERTa achieved the highest micro-average and micro-average F1 score 2 out of 3 multi-class sequence classification compared to other multilingual language models, namely mBERT and XLMR. It also yields the highest score on multi-label sequence classification. There are 2 downstream tasks for token classification task; name entity recognition (NER) and POS tagging. WangchanBERTa yields the highest score for NER but not for POS tagging. Overall, WangchanBERTa has better performance than other multilingual models.

Word embedding in social science literature

Text is a rich source of data for several fields of study. Since humans have a writing system, people record events in text, such as letters, news, and reports. Digital technology brings about a significant amount of text produced and published daily. In addition, information and knowledge are left in the text waiting to be discovered. However, text analysis requires a substantial amount of time and human work. Natural language processing makes analyzing a massive amount of text relatively quickly possible, and word embedding is an asset as it embeds the meaning within itself. The applications of word embedding in social science are illustrated in the following section.

In macroeconomics, one major concern is uncertainty: the higher uncertainty, the longer the economic downturn recovery. Kaveh-Yazdy and Zarifzadeh (2021) create an Economic Policy Uncertainty index (EPU) from word embedding. They train the Continuous Bag-of-Words Model (CBOW) on the Persian news corpus. They construct the index from the vector of words "economic", "policy", and "uncertainty". These three words are the concept they are interested. If the most similar words compared to each concept in the news have a similarity score more than the minimum threshold, that news is counted as relevant news. Next, all of the similarity scores of the relevant news are transformed into the EPU score of that news. Finally, for each month, the EPU index is calculated from the sum of the EPU score of the new in the month divided by the total news. Their EPU index performs fantastically. It follows the economic event in Iran and is aligned with the World Uncertainty Index of Iran. Their methodology has an advantage over the previously proposed method, such as LDA, or counting the news containing predefined keywords. Because the old method counts the number of news with the predefined words relevant to economic policy uncertainty (Baker et al., 2016), the incomplete relevant word set makes some news uncounted, resulting in an underperforming EPU index. This work confirms the advantages of word embedding in analyzing a large corpus to understand the topic in focus.

Word embedding is also a tool for analyzing text. In market regulation and digital market competition conduct, word embedding is used to find whether the different sizes of organizations providing digital services have a different perception of the European Digital Markets Act (DMA) and Digital Services Act (DSA). Di Porto et al. (2021) divide the text from the questionnaire on the acts into three corpora based on the size of the respondent's organization. Then, they train word embedding on each corpus and compare the distances of the words in focus between the corpora such as “Gatekeepers”, “Monopolization”, and “Newcomers”. If the vectors of the same word from different corpora are significantly different, the organizations do not have the same perception of that term. In addition, the most similar words of each group's interested terms hint at the perception of each group's concept. This is an example of how to use word embedding to find the different opinions of different groups of people utilizing the concept of word embedding similarity.

The field that leverages word embedding most is sociology, where bias and stereotype are the topics of interest. Because bias and stereotype are also within the word’s meaning, the embedding can also capture them. Without computational text analysis, concepts in sociology are studied by reading the text or survey. Therefore, studying on a large scale is limited. Garg et al. (2018) explore biases and stereotypes from the 100 years book corpus. They use the word embedding train on a decade of Google Books corpus to compute the biases. For example, they average the word vectors corresponding to female and male. Then, they calculate the distance between a vector of occupation words and the female and male vectors. If the distance from the female is higher than the distance from the male, that occupation tends to bias toward the male. They study gender bias and other biases and stereotypes and found that the bias from word embedding corresponds to the historical data. I follow the methodology of averaging the word embedding to create the representation of a concept and computing the vector similarity to compare the embedding.

The method that uses the distance of vectors is widely acceptable. Stoltz and Taylor (2021) analyze immigration from a cultural perspective by leveraging word embedding. For example, utilizing word embedding trained on decades of 100 years' book corpus, they found that the word "immigration" tends to be more similar to "crime" over time. Nelson (2021) reveals the social identity of people in the nineteenth-century U.S. South differed in gender and race from the first-person narratives.

There is also research studying the more recent events. For example, comparing the text from online platforms before and after the spread of the covid-19 pandemic, temporal word embedding reveals more Sinophobic language and insulting toward Asian people. Tahmasbi et al. (2021) train their word embedding in 3 ways. First, they train on the whole corpus. Second, they divide the text into weekly chunks. Lastly, they train the embedding on the text not in the focus period to use as the baseline. Retrieving the most similar word of the words, such as "China" and "Chinese", they found some offensive terms. There are more derogatory terms after the pandemic than in the previous period. Comparing the vocabulary of the baseline model and the vocabulary of the model train on the text of the pandemic period, they select the new terms that have a similarity score between the term and "China" of more than 0.5. Some of these are offensive. They found that the usage of these terms peaks in critical situations such as the lockdown announcement.

McCarthy et al. (2021) analyze the news of Hongkong's 2019-2020 protest from two groups of newspapers: English Language Hongkong-based newspapers and Western-based newspapers. Firstly, they train two sets of word embeddings: the Hongkong-based newspaper's embedding and the western. Then, the most similar words of protest-related keywords, such as confront, protest, and tension, are generated from each embedding. The most similar words reflect how the terms are used in the context. They found that Hongkong based language is more negative than Western-based. In addition, they split the corpus into two periods: pre-June 2019 and post-June 2019. The most similar word set of pre-June 2019 embedding and the other show diachronic semantic changes. Both groups of the newspaper report are more negative.

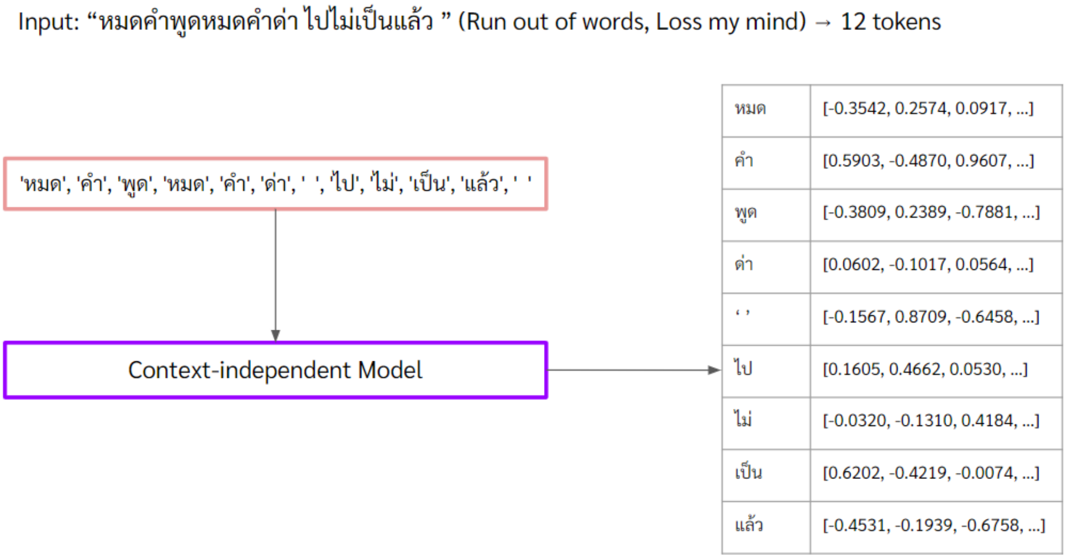
All of the works presented above use static word embedding. They trained the static word embedding on the corpus without pre-training because they wanted to analyze the idiosyncrasy within their research corpus. However, there is no work studying on the effect of pre-training on opinion analysis tasks, so I want to address this question.

Although static embedding performs well in the works, it cannot tackle polysemy because one word is represented by one fixed vector. The adoption of contextualized word embedding in social science is quite limited. Montariol et al. (2020) analyze financial text from various sources and a wide range of periods by utilizing BERT. Similar to many works presented above, they divide the text by the dimension, for instance, the time created, the author, and the industrial sector. Then, they generate the embedding. Since contextualized word embedding has one representation per occurrence, they get a set of vectors per word. They cluster the vector to find whether the word has more than one usage and how it is used in the text. Hu et al. (2019) study semantic shift through sense representation and tagging.

Hämäläinen et al. (2021) use pre-trained contextualized embeddings for depression detection in Thai text. Their work differs from other presented works as this work is a classification task. The objective is to classify whether the text contains depression content, not to analyze the content in the text. They found that LSTM with pre-trained word2vec had higher accuracy than Thai BERT. Thus, contextualized word embedding or BERT is more advantage than word2vec in classification tasks. There is a gap in whether the contextualized word embedding is still superior in the opinion analysis task.

The following figures show the difference between static word embedding and contextualized word embedding. It is able to study polysemy or do clustering from contextual word embedding because it gives a unique embedding per one instance of the word. The example text contains 12 tokens, and the static word embedding gives 9 word embeddings as 2 words occur 2 times. On the other hand, the contextualized word embedding still gives 12 embeddings.

This study will compare the two types of embeddings trained on Thai text and showcase their differences in word-based opinion analysis.



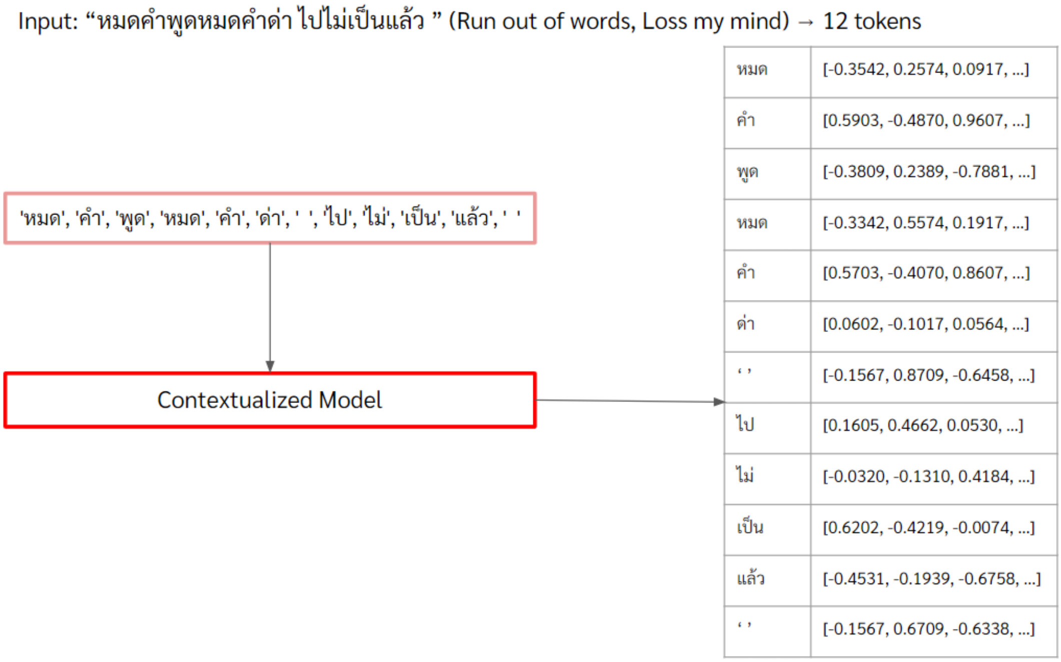
Input: “หมดคำพูดหมดคำด่า ไปไม่เป็นแล้ว ” (Run out of words, Loss my mind) → 12 tokens

'หมด', 'คำ', 'พูด', 'หมด', 'คำ', 'ด่า', ' ', 'ไป', 'ไม่', 'เป็น', 'แล้ว', ' '

Context-independent Model

หมด	[-0.3542, 0.2574, 0.0917, ...]
คำ	[0.5903, -0.4870, 0.9607, ...]
พูด	[-0.3809, 0.2389, -0.7881, ...]
ด่า	[0.0602, -0.1017, 0.0564, ...]
‘ ’	[-0.1567, 0.8709, -0.6458, ...]
ไป	[0.1605, 0.4662, 0.0530, ...]
ไม่	[-0.0320, -0.1310, 0.4184, ...]
เป็น	[0.6202, -0.4219, -0.0074, ...]
แล้ว	[-0.4531, -0.1939, -0.6758, ...]

Figure 7: Example of output from static word embedding model



Input: “หมดคำพูดหมดคำด่า ไปไม่เป็นแล้ว ” (Run out of words, Loss my mind) → 12 tokens

'หมด', 'คำ', 'พูด', 'หมด', 'คำ', 'ด่า', ' ', 'ไป', 'ไม่', 'เป็น', 'แล้ว', ' '

หมด	[-0.3542, 0.2574, 0.0917, ...]
คำ	[0.5903, -0.4870, 0.9607, ...]
พูด	[-0.3809, 0.2389, -0.7881, ...]
หมด	[-0.3342, 0.5574, 0.1917, ...]
คำ	[0.5703, -0.4070, 0.8607, ...]
ด่า	[0.0602, -0.1017, 0.0564, ...]
‘ ’	[-0.1567, 0.8709, -0.6458, ...]
ไป	[0.1605, 0.4662, 0.0530, ...]
ไม่	[-0.0320, -0.1310, 0.4184, ...]
เป็น	[0.6202, -0.4219, -0.0074, ...]
แล้ว	[-0.4531, -0.1939, -0.6758, ...]
‘ ’	[-0.1567, 0.6709, -0.6338, ...]

Figure 8: Example of output from the contextualized word embedding model

OUR APPROACH

I consider four sets of models that generate word embeddings for the focus words and their nearest neighbors. The datasets involved are 1) a large general corpus, which is typically used to pre-train language models, and 2) a focused corpus, from which I want to study the opinions. I hypothesize that how these models deal with the general corpus for pre-training, the focus corpus for fine-tuning, and the linguistic context in which word embeddings are generated will result in drastically different sets of words of all focus words.

1. Un-pre-trained static word embeddings or un-pre-trained word2vec: The utilization of un-pre-trained static word embeddings or un-pre-trained word2vec entails adopting the skip-gram model proposed by Mikolov et al. (2013) to represent static word embedding. This approach ensures that the word embeddings remain unaffected by the extensive general corpus, as they are exclusively trained on the specific focus corpus and therefore remain unprejudiced. The representation of word embeddings is derived solely from the focus corpus, enabling the identification of nearest neighbors within this embedding space for each focal word.
2. Pre-trained static word embeddings or pre-trained word2vec: The adoption of pre-trained static word embeddings or pre-trained word2vec involves the application of the skip-gram model of word2vec, whereby the model is initially pre-trained on an large general corpus and subsequently undergoes further training, commonly referred to as "fine-tuning," on the specific focus corpus for an additional epoch. In contrast to un-pre-trained embeddings, pre-trained embeddings possess prior knowledge of the Thai language acquired during the pre-training phase. This advantageous characteristic allows the embeddings to assimilate information from the focus corpus by adjusting their weights accordingly.

   I hypothesize that this set of embeddings will perform better than the un-pre-trained set when the focus corpus is small because the additional knowledge from the pre-training process on the large general corpus is required to perform well.

3. Contextualized word embeddings or WangchanBERTa: I use the model architecture from BERT (Devlin et al., 2018) to compute contextualized word embeddings. I hypothesize that contextualized word embeddings can capture the semantics of the words used in the focus corpus due to their ability to compute a word embedding from the actual context in which the word appears. For each word in the focus corpus, I find all its instances, grab the output of the last layer, and average across all instances to form a single embedding for the word. The averaged version of contextualized word embeddings retains the embedding quality and is the most effective pooling method (Bommasani et al., 2020).
4. Fine-tuned contextualized word embeddings or fine-tuned WangchanBERTa: I use the contextualized word embeddings as the same as the contextualized word embeddings, but I further fine-tuned the model on the masked language model. BERT utilizes a masked language model (MLM) to train the model to encode the language information. The intuition for the fourth type of embedding is to make the embedding embed more corpus information because the embedding gets embedding from; 1) the training on the MLM task and 2) the attention mechanism. I expect the fine-tuned contextualized word embedding to be superior to the un-fine-tuned model.

After creating these four types of word embeddings for all of the words in the focus corpus, I select focus words to analyze. A focus word is a word that represents the idea or concept that you want to study. For example, I might choose “government” as our focus word and search for the k-most similar words (nearest neighbors) to see how it is used in the focus corpus. I hypothesize that the opinion is expressed from this similar word set.

EXPERIMENT SETUP

Dataset

I use Thai protest Twitter data because it has enormous text generated by both protestors and supporters. Thus, I can analyze public opinion from this corpus. In addition, this protest has the opposition. They also have a Twitter account and generate text during the protest time so that I can compare each group's opinion without the time inconsistency.

The 2020–2022 Thai protests started in February 2020. The protesters, mostly students and young adults, first emerged from the forced dissolving of the Future Forward Party, the popular political party among the Thai younger generation. They called Prime Minister Prayuth Chan-O-Cha to resign and called for the constitution amendment. Then, the protest had more dynamic again after the reduction in the Coronavirus case. In July, big rallies started again. The demand stepped further than just the prime minister’s resignation and constitution amendment. The protest demanded monarchy reform and revoked Section 112 or lese-majesty law. ("Explainer," 2020)

Twitter is the key platform that protesters communicate, organize and discuss. Due to the decentralized nature of Twitter, it allows ordinary people to influence the movement resulting in a protest that has no clear leaders. (thaidatapointscom, 2020) Therefore, Twitter is a great source of information for studying public opinion.

I collected the data from January 12, 2020, to September 23, 2021. The keywords and hashtags to search for relevant tweets are from Mob Data Thailand (*Mob Data Thailand*, 2022), established by Amnesty International Thailand and iLaw (Thai NGO).

I retrieved the last 100 days of the period specifics above using Twitter API. 244 keywords are resulting in 343,729 tweets. Then, I clean the text by substituting a newline character (\n) with a space, removing Zero width space character (\u200b), removing the emoticon, and If hashtag characters (#) are in the middle of the tweet, remove only #, and remove both hashtag characters (#) and text otherwise. Then, I randomly selected 10 percent of the tweets for analysis, resulting in 34,186 tweets.

For the corpus of the opposition of the protester, I select the pro-monarchy accounts and their follows and their followers. Then, retrieve tweets from the same period as the protester tweets. The corpus is also cleaned by the same process as specified before, resulting in 30,106 tweets.

To prove that contextualized word embedding is more data-efficient than the static one, I sample 10 percent of the text of the corpus to create a smaller corpus. The sizes of smaller corpora are approximately 3,000 tweets. It is possible to handle by hand but is time-consuming. The datasets and their sizes are summarized in Table 1.

Name	*Corpus*	*Tweets*
The Protest Twitter	The Protest Twitter	34,186
Opposition Twitter	Opposition Twitter	30,106
The Protest Twitter - small	The Protest Twitter	3,419
Opposition Twitter - small	Opposition Twitter	3,011

Table 1: Corpus summary

Embeddings

The un-pre-trained static word embedding (Figure 9) utilized the skip-gram model of word2vec (Mikolov et al., 2013). The embedding considers 5 words window and has 300 dimensions. The embeddings are initialized randomly and then fine-tuned on the focus corpus. I use the gensim[1] implementation of word2vec.

The pre-trained static word embeddings (Figure 10) are pre-trained on open-access Thai text (Lowphansirikul et al., 2021). In total, these datasets are 4.33 GB. I use the skip-gram model of word2vec (Mikolov et al., 2013) for static word embedding. The model uses a 5-word window and 300 dimensions, similar to the un-pre-trained static word embedding. The model is pre-trained for one epoch. Then, I fine-tune the embeddings on the focus corpus one epoch to adjust the weight to acquire more specific knowledge.

[1] http://radimrehurek.com/gensim

I use the contextualized word embeddings (Figure 11) from WangchanBERTa (Lowphansirikul et al., 2021). In addition to the data used to pre-trained the static word embeddings, this model was also pre-trained on social media data (wisesight-large), web forum data from Pantip.com (pantip-large), and Thai National Corpus (TNC) (Aroonmanakun et al., 2009). In total, the dataset of WangchanBERTa is 78.5 GB. WangchanBERTa uses the objective function of RoBERTa (Liu et al., 2019). The model takes a sequence of 512 tokens, and the output layer has 768 dimensions.

Lastly, I use the WangchanBERTa with the same setting as the contextualized model for fine-tuned contextualized word embedding (Figure 12). I additionally trained the WangchanBERTa on the masked language model task on the focus corpus for 3 epochs. Then, the final model is used to get the word embedding.

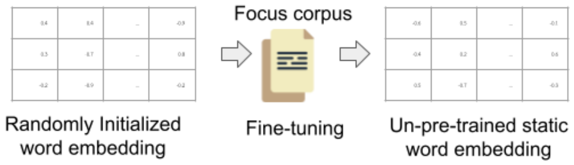

Focus corpus
Randomly Initialized word embedding
Fine-tuning
Un-pre-trained static word embedding

Figure 9: Un-pretrained static word embedding

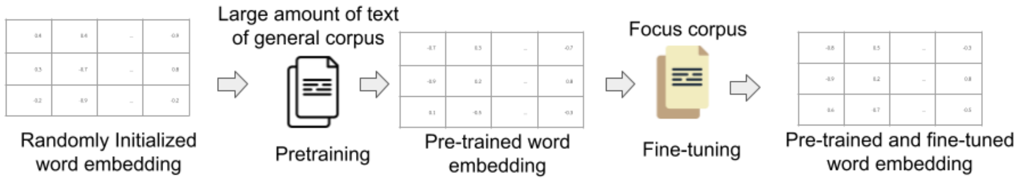

Large amount of text of general corpus
Focus corpus
Randomly Initialized word embedding
Pretraining
Pre-trained word embedding
Fine-tuning
Pre-trained and fine-tuned word embedding

Figure 10: Pre-trained static word embedding

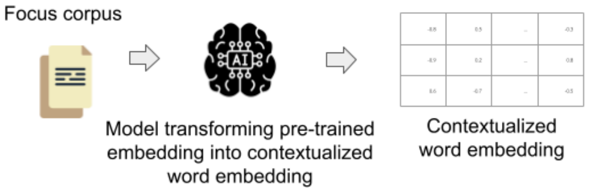

Focus corpus
AI
Model transforming pre-trained embedding into contextualized word embedding
Contextualized word embedding

Figure 11: Contextualized word embedding

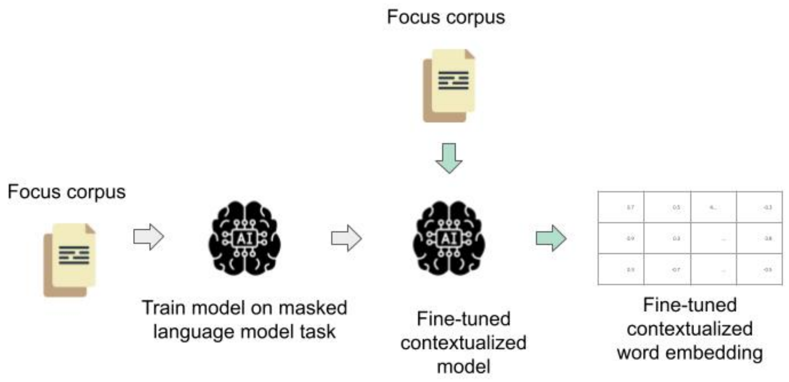

Focus corpus
Focus corpus
Train model on masked language model task
Fine-tuned contextualized model
Fine-tuned contextualized word embedding

Figure 12: Fine-tuned contextualized word embedding

After get the embedding of each type of model, I get the top 10 most similar words of our focus words for the analysis. The most similar word is the word that has the highest cosine similarity when comparing it with the focus word. The formula for cosine similarity is defined in Equation 16 and the process is shown in Figure 13.

$$Sim(V_x, V_y) = Cosine(\theta) = \frac{V_x \cdot V_y}{\|V_x\|\|V_y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt[2]{\sum_{i=1}^{n} x_i}\ \sqrt[2]{\sum_{i=1}^{n} y_i}}$$

Equation 16

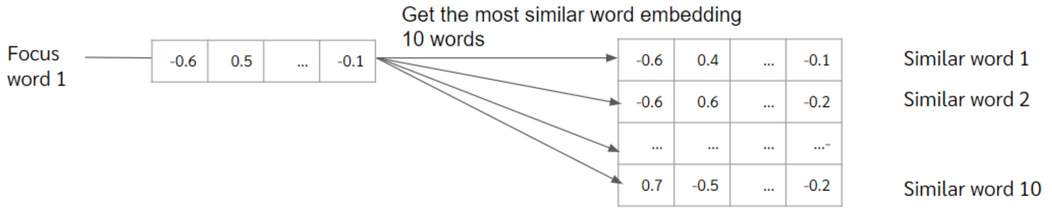


Figure 13: The process to get top 10 most similar word

Clustering

As contextualized word embedding has one unique embedding per occurrence, it is possible to cluster these embeddings. I hypothesize that the uses of a word in opposing corpora are different. Therefore, I could observe the two clusters of embeddings.

The algorithm for clustering in this study is KMeans. This algorithm tries to assign the data points into k groups by considering the distance between the data point and the cluster's centroid. The number of clusters or k is predefined before running. Kmeans works in the following steps.

1. The algorithm randomly picks the centroid of each cluster. For example, if the k is equal to 4, there are 4 centroids.
2. The algorithm calculates the distance between each data point and the centroids. I chose the cosine distance as a metric to make this align with the process for similar word retrieving in the previous step. The data point will be assigned to a cluster with the minimum distance.
3. The algorithm will re-initialize the centroid by averaging all data points within the cluster.
4. The algorithm will repeat steps 2 and 3 until the data points do not change the clusters they belong.

(Jin & Han, 2010)

Since the number of k must be pre-defined, I use the elbow method to select the optimal k number. The elbow method runs the KMeans clustering with the number of k starting from 1 to a reasonable number of groups. For instance, considering the optimum number of k ranging from 2 to 10, the algorithm will run the KMeans and calculate the average square distances between the data points to the calculated centroid (distortions) of each k. Then, the average square distances are plotted. The optimum k is the value at the curve suddenly decreases, called elbow point. ("Elbow Method for optimal value of k in KMeans," 2019) The visualization of this method is presented in Figure 14.

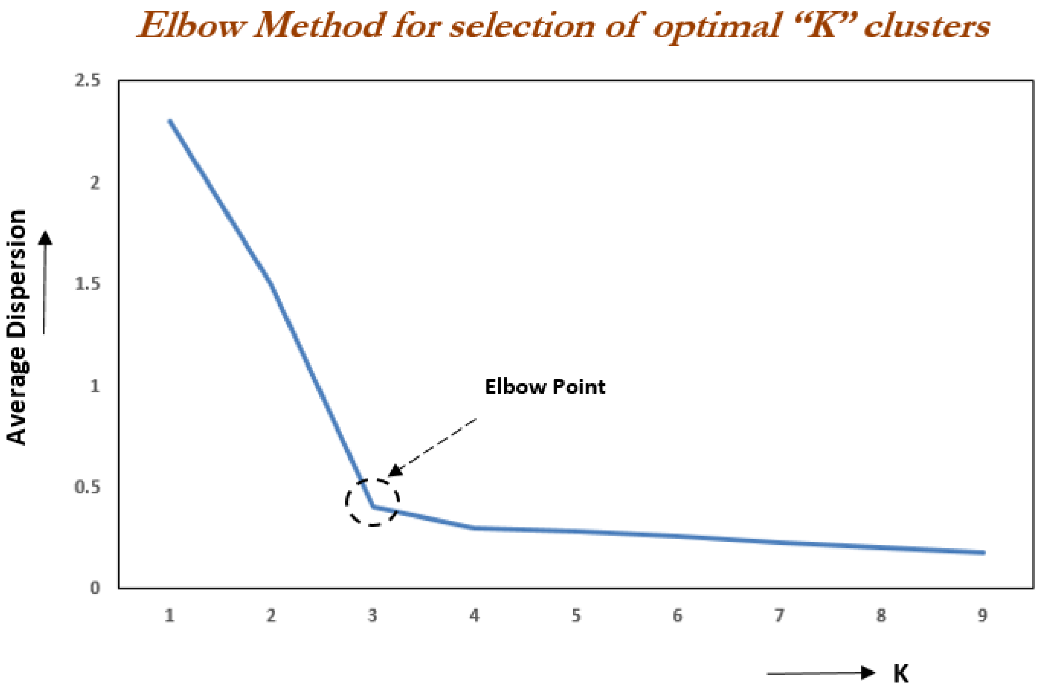


Figure 14: Example of elbow method (The Elbow Method - Statistics for Machine Learning [Book], n.d.)

Evaluation

I evaluate the quality of the similar set given a focus word by three measures: the proportion of the words that relate to their focus words syntactically, the proportion of the words that relate to their focus words semantically, and the proportion of the words that are useful for the word-based opinion analysis.

I label each word in a similar set as syntactically related to the focus words if it is in the same syntactic tree as the focus word. Similarly, I label each word in a similar set as semantically related to the focus word if the two words hold one of the lexical relations drawn from Wordnet[2], e.g., troponym, antonym, coordinate, entailment, holonym, hypernym, hyponym, meronym, synonym. I use the relations from Wordnet as an idea to assign the relation.

Note that a word in a similar set might have both syntactic and semantic relation to the focus words, or it might have no relation to the focus word.

[2] https://wordnet.princeton.edu/documentation/wngloss7wn

In the context of analyzing opposing opinions, “useful” words are defined as those that demonstrate corpus-specific opinions that contrast or conflict with other corpora. A useful word in a similar set should reflect the opinion expressed in the corpus. For example, if the focus word is “protest,” words like “gather” or “assembly” are not useful for describing opposing opinions because they reflect shared opinions or lexical relations to the focus word. Instead, a word like ป่วน (“be turbulent”) from the opposition corpus would show a negative perspective toward the protest.

I perform the Chi-square test for all of the following analysis to ensure that the differences between the types of embedding are statistically significant.

RESULTS AND DISCUSSION

I select 9 words to analyze the result in detail, similar to the previous works in word-based analysis in social sciences. For example, Tahmasbi et al. (2021) studied racism against Chinese people; their focus words are “china” and “Chinese”. Because this study has an objective to do the opposing opinion analysis, I choose 9 focus words that represent the topic of interest of the political protest. These 9 words represent the topic of interest and are categorized into 4 topics.

The first topic is the protest itself. Its illustrated words are แกง “deceive”, ม็อบ “protest”, and เรียกร้อง “demand/call for”. The word แกง “deceive” originally meant “curry” or “make a curry”. It also means “deceive”, first used in the homosexual community. Then, the alternative became popular among the protesters. This word is one of the various slang that emerged during the events. (Thairath, 2020) We select this term because it expresses how the protesters and the opposition use the language differently. Regarding เรียกร้อง “demand/call for”, the protesters declare key demands that the government must fulfill. Therefore, it is worthwhile to analyze the opinion surrounding the demands.

The second topic is the opinion of the protesters toward the opposition. We selected the term สลิ่ม “Sa-lim”, which is the name that the protesters call the opposition or the conservatism in derogatory ways. (Thairath, 2021)

The third topic is the concept of nation; the focus words consist of ชาติ “nation”, ประชาชน “people”, and ประเทศ “country”. The importance of this topic comes from the argument from the opposition. They called the protesters the haters of the nations (ชังชาติ) or the haters of the homeland. I translate the word ชาติ according to the translation in most dictionaries. However, the word ชาติ or ความรักชาติ “nationalism” for the opposition is close to the sense of patriotic or the love of homeland instead of nationalism, while the protesters used ชาติ “nation” in the sense of nation. Therefore, one of the main conflicts between these opposing groups is due to the different in the sense of the word ชาติ “nation”. (Kam-Phaka, 2019) In addition, the word ชาติ in Thai usually occur with the word ประเทศ “country” to form a phrase, and sometimes the word ประเทศ “country” is used in the sense of nation. Thus, both words should be analyzed. The word ประชาชน “people” is important because people are the main composition of a country, and the argument is political protest is usually about the entity in the country, so it is worthwhile to explore this focus word.

Lastly, the topic of political policy and entity contains ภาษี “tax” and รัฐบาล “government”. As the protesters shouted “Our tax” and “The most delicious thing is the people’s tax” during the demonstration, these phrases became viral (iLaw, 2021). I select this word due to its significance. In addition, both protesters and the oppositions call the government to do or not to do somethings, so the opposing opinions in the word รัฐบาล “government” should be considered.

After I have selected all focus words, I recheck the significance of each word from the frequency in the dataset. Most of the focus words have a high frequency. The percentiles of the frequencies are greater than 90 except for the word แกง “deceive” in the opposition's small-size dataset (82.02 percentile). Thus, all of them can represent their opinions in the corpora. All words and their percentile are presented in Table 2.

Focus word	Protestor		Opposition	
	Small size dataset	Large size dataset	Small size dataset	Large size dataset
แกง “deceive”	94.33	98.62	82.02	96.75
ชาติ “nation”	96.46	99.15	98.07	99.46
ประชาชน “people”	99.62	99.91	96.60	99.37
ประเทศ “country”	99.05	99.77	98.98	99.77
ภาษี “tax”	96.78	99.31	94.06	98.77
ม็อบ “protest”	99.38	99.87	97.91	99.58
รัฐบาล “government”	98.35	99.64	98.29	99.58
เรียกร้อง “demand/call for”	97.42	99.36	90.66	98.84
สลิ่ม “Sa-lim”	96.94	99.29	97.64	99.51

Table 2: The percentile of the frequency of 9 selected focus word

I show similar word sets of 3 focus words in the following tables.

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
รอบ "turn, around"	สิทธิ์ "right"	ประณีประนอม "compromising"	ผิดหรอ “Is wrong?”	ประท้วง "protest"	ประท้วง "protest"	ประท้วง "protest"	ประท้วง "protest"
น. "the abbreviation that can refer to several words"	พื้นฐาน "fundamental"	ผินหลังให้ "ignore"	ทั้งๆที่ “in spite of”	ต้องการ "require"	ทวง "demand"	ขอความ "require"	ต่อสู้ "fight"
เอง "by oneself"	แค่ "only"	ผ่อนสั้นผ่อนยาว "be flexible, compromise"	มึง “you”	ต่อสู้ "fight"	ต่อต้าน "resist/against"	ต้องการ "require"	ต่อต้าน "resist/against"
สิ่ง "thing"	สิ่ง "thing"	สำรวมใจ "calm"	ทำร้ายปชช “attack the people”	ขอความ "require"	ต่อสู้ "fight"	ต่อสู้ "fight"	ทวง "demand"
เล่น "play"	เป็นธรรม "be fair"	ศักเคียส "not a word, unable to translate"	กิบลัต “not a word, unable to translate”	ร้องขอ "ask for request, demand"	ร้องขอ "ask for request, demand"	ร้องขอ "ask for request, demand"	ทวงคืน "revoke"
แตก "split, break"	ประชาธิปไตย "democracy"	ความร้อนรน "impatience"	อิกนอร์“Ignore”	ปกป้อง "protect"	ทวงคืน "revoke"	สู้เพื่อ "fight for"	อยากได้ "want"
เด็ก "young people"	ถูกต้อง "right"	medicare	กินดี“part of the phrase live happily”	สู้เพื่อ "fight for"	ขับไล่ "drive out"	ปกป้อง "protect"	แสดงออก "express, demonstrate"
หน่วย "division, unit"	ด้วยซ้ำ "as well"	ayah	แหล่ะ “ (end) particle used at the end of a statement to emphasize it”	ชุมนุม "gather"	แสดงจุดยืน "stand up for"	ชุมนุม "gather"	ปกป้อง "protect"
ด่าน "barrier"	รับฟัง "listen to"	ทุ่มเถียง "argue"	โอนอ่อนผ่อนตาม “give in”	อยากได้ "want"	แสดงออก "express, demonstrate"	เรียก "call"	ต้องการ "require"
กลุ่ม "group"	ผิดหรอ "Is it wrong?"	เข้าร่องเข้ารอย "act in the right way"	” “	เรียก "call"	สู้เพื่อ "fight for"	แสดงความ "show"	สู้เพื่อ "fight for"

Table 3: Similar word set of the focus word เรียกร้อง "demand/call for" from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ลด "reduce"	สิทธิ "right"	ประณีประนอม "compromising"	บุลลี่ “bully”	แอนตี้ "anti"	ประท้วง "protest"	ประท้วง "protest"	ประท้วง "protest"
ภูมิ "background"	อ้าง "refer to, claim"	ผินหลังให้ "ignore"	ปชช “abbreviation of people”	เสรีภาพ "freedom"	ต่อต้าน "resist/against"	เสรีภาพ "freedom"	ทวง "demand"
ปล่อย "allow, to release"	ละเมิด "infringe"	ผ่อนสั้นผ่อนยาว "be flexible, compromise"	บุลลี่ “bully”	นม "breast, milk"	ต้องการ "require"	โหน "hang"	ต่อต้าน "resist/against"
เห็น "see"	กระบวนการ "process"	ความร้อนรน "impatience"	สามกีบ "three-toes hoof"	เห็นต่าง "disagree"	ทวง "demand"	ทวง "demand"	กดดัน "force"
เข้า "enter"	โจมตี "attack, assault"	สำรวมใจ "calm"	หลับหูหลับ “part of the phrase meaning blindly”	ขยัน "hard-working"	ถามหา "ask"	เคลม "claim"	ถามหา "ask"
เสียใจ "be sorry, regret"	เสรีภาพ "freedom"	พสก "people"	กิบลัต “not a word, unable to translate”	เสียงดัง "loud"	ร้องขอ "ask for request, demand"	ปะทะ "crash"	รณรงค์ "campaign"
ใหม่ "new"	ต่อต้าน "resist/against"	ใฝ่ต่ำ "behave badly"	มาดราม่า “be dramatic”	หลุด "slip out"	กดดัน "force"	อยากได้ "want"	ต้องการ "require"
เดิม "at first"	มนุษยชน "human"	ทุ่มเถียง "argue"	กฎหมาย “law”	ร้านค้า "store"	รณรงค์ "campaign"	ประกาศ "announce"	สนับสนุน "support"
เฉพาะ "be specific"	ยัดเยียด "force"	ศักเคียส "unable to translate"	บังคับใจ “control one’s mind”	io "information operation"	โหยหา "yearn for"	งอแง "be petulant"	ขับไล่ "drive out"
รถ "vehicle"	กฎหมาย "law"	คาดหมาย "expect"	” “	ประท้วง "protest"	โวยวาย "outcry"	ต้องการ "require"	งอแง "be petulant"

Table 4: Similar word set of the focus word เรียกร้อง "demand/call for" from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
สิ่ง "thing - usually used as the rightful thing in the corpus"	หม้อ "pot"	ไตปลา "fish's kidney"	ไตปลา "fish's kidney"	ฉีดน้ำ "spray water"	พยายามฆ่า "make an attempt on someone's life"	ฉีดน้ำ "spray water"	จัดฉาก "set the scene"
แม่ "mother"	ยูธ "subword of the group’s name, free youth"	ฮังเล "Hunglei curry"	ฮังเล "Hunglei curry"	ฉีดวัคซีน "inject vaccine"	บูด "rancid - it can be used as slang meaning bad"	ทุบ "beat"	ร้องเรียน "complain"
โลก "world"	แผน "plan"	กะหรี่ "curry"	ส้มกุ้ง "Som Khung - a Thai dish"	ขับรถชน "crash car"	จัดฉาก "set the scene"	ฉีดวัคซีน "inject vaccine"	สกัด "intercept"
น่า "should"	แกนนำ "leader, mainstay"	ส้มกุ้ง "Som Khung - a Thai dish"	เขียวหวาน "green curry"	สกัด "intercept"	แก๊ส "gas"	แก๊ส "gas"	ซ่อม "repair"
ค่า "fee"	แก "you"	เขียวหวาน "green curry"	น้ำแกง "soup"	ล้อม "enclose"	โจมตี "attack"	ล้อม "enclose"	แก๊ส "gas"
สามารถ "able to"	โอเค "OK"	เครื่องแกง "spices"	พริกขิง “Phrik Khing, one of Thai dish”	ขยะแขยง "disgusting"	ทำงานให้ "work for"	เดือด "be angry/boil"	โจมตี "attack"
น. "the abbreviation that can refer to several words"	รับผิดชอบ "be responsible"	น้ำแกง "soup"	น้ำเงี้ยว "Nam Ngiao - a Thai dish"	แก๊ส "gas"	สกัด "intercept"	ซ่อม "repair"	ทุบ "beat"
ตั้งแต่ "since"	เตือน "remind, warn"	แกงเขียวหวาน "green curry"	บุ๊ง “unable to translate”	บูด "rancid - it can be used as slang meaning bad"	ต้อน "herd"	เครื่องแบบ "uniform"	ฉีดน้ำ "spray water"
รอบ "turn, around"	สถานที่ "location"	น้ำพริก "chili paste"	ยูธ “subword of the group’s name, free youth”	อห "the abbreviation of swearword"	ทิพย์ "unreal"	สกัด "intercept"	พระราชทาน "bestow on"
...	จัด "arrange"	น้ำเงี้ยว "Nam Ngiao - a Thai dish"	น้ำพริก "chili paste"	ไหม้ "burn"	ไปแจ้งความ "report, inform, notify to the polices"	จงใจ "intend"	บูด "rancid - it can be used as slang meaning bad"

Table 5: Similar word set of the focus word แกง "deceive" from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ผม "hair, I"	เจียว "render"	ไตปลา "fish's kidney"	ไตปลา "fish's kidney"	ข้าว "rice"	ต้ม "boil"	ข้าว "rice"	ต้ม "boil"
ตรง "straight"	หอม "shallot"	ฮังเล "Hunglei curry"	ฮังเล "Hunglei curry"	ผัก "vegetable"	แกงส้ม "sour soup made of tamarind paste"	อาหาร "food"	ปลาร้า "pickled fish"
out	ทอด "fry"	ส้มกุ้ง "Som Khung - a Thai dish"	ส้มกุ้ง "Som Khung - a Thai dish"	อาหาร "food"	ข้าวต้ม "soft-boiled rice"	ผัก "vegetable"	แกงส้ม "sour soup made of tamarind paste"
ชวน "persuade"	หม้อ "pot"	เขียวหวาน "green curry"	เขียวหวาน "green curry"	หมี่ "rice noodle"	น้ำพริก "chili paste"	ต้ม "boil"	น้ำพริก "chili paste"
เข้า "enter"	😋	กะหรี่ "curry"	อ่อม "Aom - a kind of Thai curry"	ก๋วยเตี๋ยว "noodle"	ก๋วยเตี๋ยว "noodle"	น้ำ "water"	สุกี้ "Sukiyaki"
ชาติ "nation"	หั่น "cut"	อ่อม "Aom - a kind of Thai curry"	น้ำแกง "soup"	ต้ม "boil"	ปลาทู "mackerel"	ก๋วยเตี๋ยว "noodle"	ยำ "Yum - a Thai dish"
หน่อย "slightly"	ปลา "fish"	น้ำแกง "soup"	น้ำเงี้ยว "Nam Ngiao - a Thai dish"	น้ำ "water"	ไข่ "egg"	หมี่ "rice noodle"	ไข่เจียว "omelet"
ใส่ "put something in"	ยยยย "a cluster of a charactor"	แกงเขียวหวาน "green curry"	พริกขิง “Phrik Khing, one of Thai dish”	ขิง "ginger"	ยำ "Yum - a Thai dish"	หมู "pork"	ไข่ "egg"
เหตุ "reason"	นม "breast, milk"	เครื่องแกง "spices"	พะแนง “dry curry”	หมู "pork"	กะทิ "coconut milk"	บัตร "card, ticket"	ผัด "fry"
ข้าง "side"	ไก่ "chicken"	น้ำพริก "chili paste"	เคร่อื งแกง “spices”	เย็น "cool"	ผัด "fry"	ขิง "ginger"	กะทิ "coconut milk"

Table 6: Similar word set of the focus word แกง "deceive" from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
วัน "day"	ธรรมเนียบรัฐบาล "Government House"	ม็อป "mob, protest"	ม๊อบ "mob, protest"	ม๊อบ "mob, protest"	ม๊อบ "mob, protest"	ม๊อบ "mob, protest"	ชุมนุม "gather"
เปิด "open"	วัน "day"	vileda (name of mop's brand)	ม็อป "mob, protest"	ผู้ชุมนุม "protester"	ชุมนุม "gather"	ชุมนุม "gather"	ม๊อบ "mob, protest"
ฝาก "leave"	,rt	ม๊อบ "mob, protest"	ม็อบ26 "mob, protest on 26th"	ชุมนุม "gather"	การ์ด "the volunteer guard"	การชุมนุม "assembly"	ประท้วง "protest"
ล้าน "million"	Please	newsponge (name of mop's brand)	ม้อบ "mob, protest"	การชุมนุม "assembly"	คาร์ "car"	ผู้ชุมนุม "protester"	แกนนำ "leader, mainstay"
)		pva (a type of cleaning sponge)	newsponge (name of mop's brand)	การ์ด "the volunteer guard"	แกนนำ "leader, mainstay"	แกนนำ "leader, mainstay"	การชุมนุม "assembly"
เจอ "find"	สาธารณะสุข "public health"	swash (name of mop's brand)	Retweet	ล้อม "enclose"	ผู้ชุมนุม "protester"	แยก "side road"	สถานการณ์ "situation"
เพิ่ม "add"	Retweet	supercat (name of mop's brand)	ม้อบ "mob, protest"	แกนนำ "leader, mainstay"	หาดใหญ่ "Hat Yai"	การ์ด "the volunteer guard"	ขบวน "procession"
หน่วย "division"	หาดใหญ่ "Hat Yai"	ไม้ถูพื้น "mop"	supercat (name of mop's brand)	รัฐประหาร "coup d'etat"	มวลชน "masses"	ประท้วง "protest"	ไฟไหม้ "be on fire"
แกนนำ "leader, mainstay"	ธรรมเนียบ "residence"	be-man	ธรรมเนียบ "residence"	แยก "side road"	อาชีวะ "vocational students"	โจร "robber"	การ์ด "the volunteer guard"
เรา "we"	💢	ธรรมเนียบ "residence"	pva (a type of cleaning sponge)	แก๊สน้ำตา "tear gas"	การจับกุม "arresting"	ล้อม "enclose"	แยก "side road"

Table 7: Similar word set of the focus word ม็อบ "protest" from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
จาก "from"	ม๊อบ "mob, protest"	ม็อป "mob, protest"	ม็อป "mob, protest"	ม๊อบ "mob, protest"	ม๊อบ "mob, protest"	ม๊อบ "mob, protest"	ชุมนุม "gather"
ขอ "beg"	ชุมนุม "gather"	vileda (name of mop’s brand)	ม๊อบ "mob, protest"	ชุมนุม "gather"	ชุมนุม "gather"	ชุมนุม "gather"	ม๊อบ "mob, protest"
ผล "result"	ป่วน "be turbulent"	ม๊อบ "mob, protest"	vileda (name of mop’s brand)	แกนนำ "leader, mainstay"	แกนนำ "leader, mainstay"	แกนนำ "leader, mainstay"	แกนนำ "leader, mainstay"
สิ่ง "thing"	สลาย "disband"	newsponge (name of mop’s brand)	newsponge (name of mop’s brand)	การชุมนุม "assembly"	การชุมนุม "assembly"	การชุมนุม "assembly"	การชุมนุม "assembly"
ขนาด `size"	ตำรวจ "police"	pva (a type of cleaning sponge)	pva (a type of cleaning sponge)	ประท้วง "protest"	ประท้วง "protest"	ประท้วง "protest"	ประท้วง "protest"
รถ "vehicle"	เผา "burn"	ไม้ถูพื้น "mop"	supercat (name of mop’s brand)	ขบวนการ "movement"	ผู้ชุมนุม "protester"	สลิ่ม	เสื้อแดง "Red shirt"
โดย "by"	ฝูง "herd"	supercat (name of mop’s brand)	สามกีบ "three-toes hoof"	ป่วน "be turbulent"	เสื้อแดง "Red shirt"	ตำรวจ	แก๊ง "Gangs"
2	ดินแดง "Din-Dang"	swash (name of mop’s brand)	ไอ้เหี้ย “Goddammit!”	สลิ่ม "Sa-lim"	แก๊ง "Gangs"	ผู้ชุมนุม "protester"	ผู้ชุมนุม "protester"
อะ "a"	คฝ "Crowds Control polices"	be-man	swash (name of mop’s brand)	มวลชน "masses"	เด็กแว้น "motorpunk"	มวลชน	แท็ก "tag"
ยุค "age"	แกนนำ "leader, mainstay"	ncl	ประนาม “condemn”	ตำรวจ "police"	ฝูงชน "crowd"	ป่วน	ฝูงชน "crowd"

Table 8: Similar word set of the focus word ม็อบ “protest” from opposition corpus

1. How does pre-training affect the output of the static word embeddings in the word-based opinion analysis task?

To answer this question, I analyze the effect of pre-training in 2 criteria: the quality of word embedding and the effect on word-based opinion analysis tasks.

This investigation examines the impact of pre-training by conducting a comparative analysis between un-pre-trained word2vec and pre-trained word2vec models.

In this study, embedding quality is evaluated based on the assessment of the proportion of words exhibiting either semantic or syntactic relations that are similar to the focus word. A high-quality embedding is expected to yield the most similar words that possess a meaningful connection to the focus word, either through semantic relation or through co-occurrence within the same syntactic tree. Conversely, if the word embedding fails to retrieve words that demonstrate these relations, it indicates a low-quality embedding.

Based on the analysis of the similar word sets, it has been observed that when the size of a focused corpus is small, the sets of similar words associated with the focus words tend to lack interpretability. However, the pre-training process has positively impacted the overall quality of the embeddings. This study quantifies the number of words that demonstrate syntactic or semantic associations in a general contextual sense to assess the relevance of each similar set concerning their respective focus words.

In the case of a dataset characterized by a small sample size, pre-training of word embeddings yields superior quality, as evidenced by the larger number of similar words exhibiting meaningful relations within the pre-trained model. This improvement is supported by the findings in Figure 15, where the proportion of similar words lacking discernible relation decreased significantly from 65.0% to 36.7% (χ2 test; $p < 0.05$).

An example of the improved word quality is shown in. for the focus word เรียกร้อง “demand/call for”, the top 3 most similar words of the pre-trained model are ประณี-ประนอม “compromising”, ผินหลังให้ “ignore”, ผ่อนสั้นผ่อนยาว “be flexible, compromise”. On the other hand, the set of un-pre-trained embedding contains unrelated meaning words.

Conversely, the effectiveness of pre-training in enhancing embedding quality is not observed in the case of a large corpus. As depicted in Figure 16, the proportion of similar words lacking any discernible relation increased from 25.6% to 41.1% following pre-training (χ2 test; $p < 0.05$). This result suggests that using pre-training in the context of a large corpus does not yield the desired improvements in embedding quality.

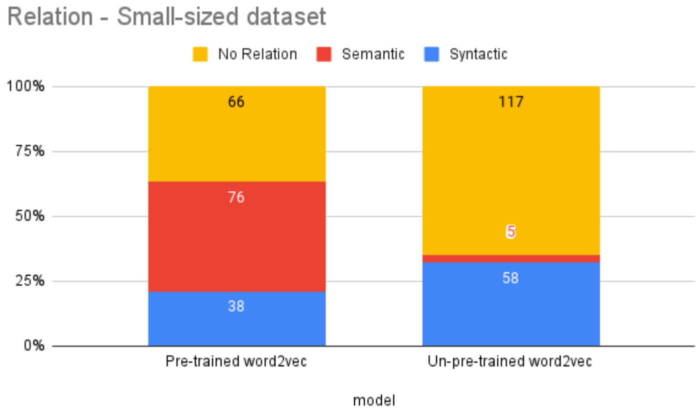


Figure 15: Proportion of similar words with relations and no relation for un-pre-trained and pre-trained word2vec of the small-sized dataset

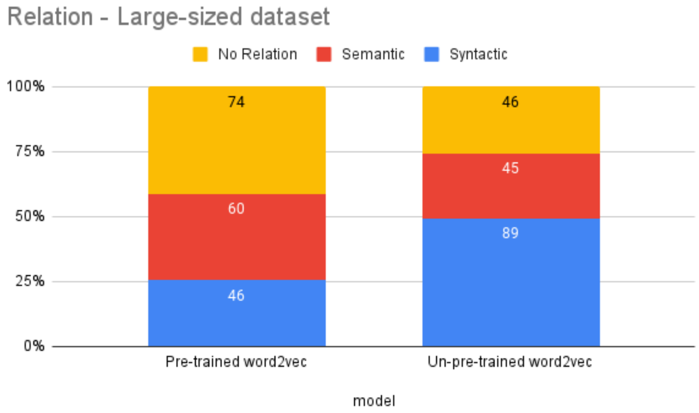


Figure 16: Proportion of similar words with relations and no relation for un-pre-trained and pre-trained word2vec of the large dataset

In word-based opinion analysis, the number of words expressing specific information within the small dataset remains significantly low regardless of whether an un-pre-trained or pre-trained model is employed. As depicted in Figure 17, the proportion of words deemed useful stands at 3.3% for the un-pre-trained model and 2.8% for the pre-trained model (χ2 test; $p \geq 0.05$). While the pre-training process enhances the overall quality of word embeddings, it does not effectively aid in extracting valuable, corpus-specific opinions. Consequently, when dealing with an excessively small focused corpus, a static word embedding technique such as word2vec proves ineffective in the context of word-based opinion analysis, irrespective of the incorporation of the pre-training process.

Here are some examples of useful words in the small set of the un-pre-trained word2vec. The focus word ประเทศ “country” of the protesters has the useful word ปัญหา “problem”. This word is useful because it shows the topic associated with the focus word. The protestors see that there are problems in the country. The other example is the focus word ภาษี “tax” has the useful word กษัตริย์ “monarch”. It shows debates about the tax allocated to the royal institution. However, these useful words still show a shallow analysis and just show the topics related to the focus words. There is only one useful word from the opposition corpus, i.e., ประเทศ “country” and its similar word ทำลาย “destroy”. Though both protesters and oppositions have negative opinions on the focus word ประเทศ “country”, oppositions see someone or an enemy destroying the country. In contrast, the protesters focus on the problems themselves.

The number of useful words of the pre-trained word2vec is as small as that of the un-pre-trained word2vec. The protesters’ focus word รัฐบาล “government” have similar words ทำร้ายปชช “attack the people”, and เล่นสกปรก “play a dirty trick”. They show a negative opinion toward the government, and the protesters distrust it. On the other hand, the useful word of the opposition is เฟคนิวส์ “fake news”. It is quite difficult to analyze, but when considering the text samples, the oppositions see that the government is attacked by fake news.

Compared to a small number of useful words in the small-sized datasets, the number of useless similar words is substantially larger. The useless words usually have a general sense of the word instead of the corpus-specific senses. The salient example is the set of focus word แกง “deceive” (Table 5). In general Thai, this word originally means curry (n.), but it means to deceive in the protesters’ corpus context. In the protester corpus, the pre-trained embeddings yield only words related to curry types or food. This situation occurs for other focus words.

Out of 9 focus words, the word-based analysis can be done on just 1 to 2 focus words with only less than half of all similar word. Thus, this task is not recommended when the training set is insufficient.

Moreover, the effectiveness of pre-training for word-based opinion analysis in a large-sized dataset is also limited. Figure 18 illustrates that the proportion of useful words declines significantly from 62.8% to 26.1% following the application of pre-training (χ2 test; $p < 0.05$). These findings suggest that pre-training does not yield favorable outcomes in enhancing the extraction of valuable words for opinion analysis in a large dataset.

There is an example in the case of the focus word ม็อบ “protest”. In the Thai language, “mob” is used in the sense of protest and gathering together without the negative connotation. This word is transliterated in Thai, and its spelling is homograph and homophone of the transliterated word “mop”. In Table 7, the sets of the pre-trained models contain several mop brands, e.g., vileda, newsponge, supercat, pva. The other similar words are the word meaning protest in varied spelling that does not express any corpus-specific meaning. On the other hand, a similar set of the un-pre-trained models does not contain any words related to cleaning applications. This means the un-pre-trained model is better at extracting corpus-specific meaning.

Therefore, when dealing with a sufficiently large focused corpus, it is observed that similar sets derived from un-pre-trained word embeddings yield more valuable and corpus-specific outcomes compared to the sets derived from pre-trained word embeddings.

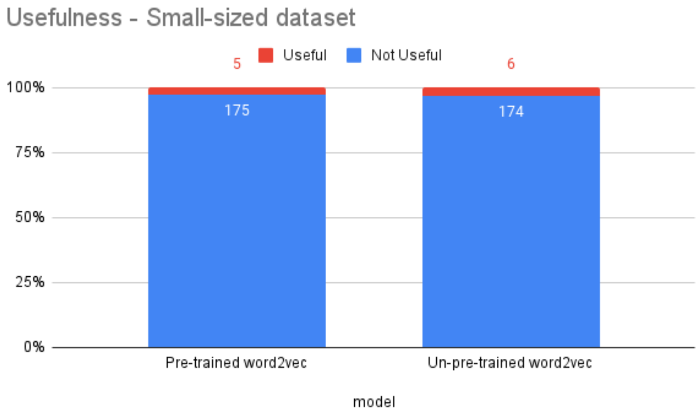


Figure 17: Proportion of useful words from a small-sized dataset

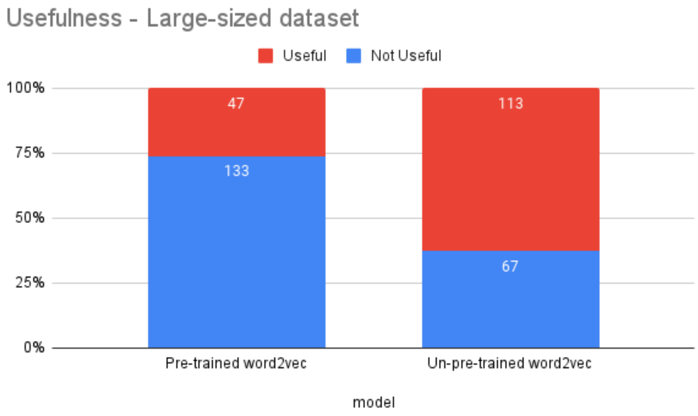


Figure 18: Proportion of useful words from a large-sized dataset

This outcome can be attributed to a significantly large open-access Thai text corpus (Lowphansirikul et al., 2021), which encompasses a size of 4.33 GB. Pre-training enhances the quality of word embeddings within small-sized word2vec models, as the focus corpora in these cases are too limited to effectively adjust the embedding weights from random initialization to their optimal values. However, the aforementioned Thai text corpus, being considerably larger in scale compared to the focus corpora under investigation, presents challenges when further training the pre-trained Word2Vec model. This could result in minor weight adjustments or introduce noise into the embeddings. Consequently, it is advisable to refrain from employing pre-training techniques for static word embeddings for large datasets.

2. Which of the two embeddings is more useful in delineating the opposing opinion in two opposing focus corpora?

Based on the findings presented in Figure 19, the un-pre-trained word2vec model demonstrates the highest count of useful words, with 119 out of 360 words (33.06%). The contextualized model, specifically WangchanBERTa, follows closely with 97 out of 360 words (26.94%) identified as useful. The fine-tuned WangchanBERTa model ranks third in usefulness, with 80 words (22.22%). Conversely, the pre-trained word2vec model exhibits the lowest number of useful words, with only 52 words (14.44%) identified as valuable ($\chi 2$ test; $p < 0.05$). These findings indicate that the un-pre-trained word2vec model is the most effective for distinguishing between two contrasting corpora. In contrast, the pre-trained word2vec model is not recommended for word-based opinion analysis tasks.

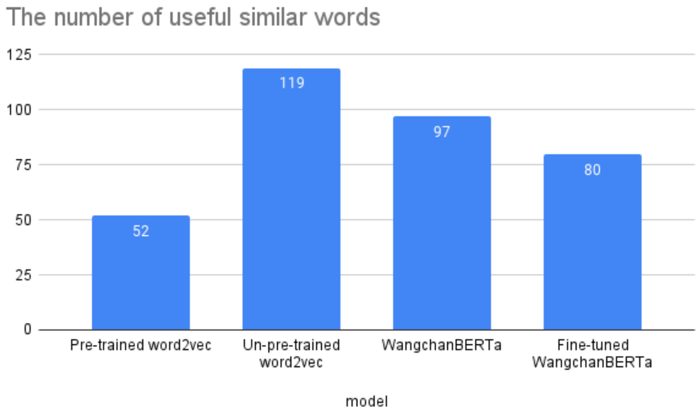


Figure 19: Number of useful similar words in delineating opposite opinions of each model

Nonetheless, when considering the influence of dataset size, it is evident from Figure 20 that static word embeddings demonstrate limited effectiveness in the context of word-based opinion analysis on a small corpus. The pre-trained and un-pre-trained models yield minimal useful words, constituting only 2.78% and 3.33%, respectively. In contrast, the contextualized models exhibit superiority in this regard. Specifically, WangchanBERTa captures some opposing opinions in 24.44% of the similar words (44 out of 180), followed by fine-tuned WangchanBERTa at 22.22%. These findings underscore the advantages of employing contextualized models over static word embeddings when dealing with smaller corpora in word-based opinion analysis tasks.

With a larger corpus, un-pre-trained word2vec outperforms others, with 113 useful words out of 180 (62.78%). WangchanBERTa has 53 (29. 44 %), and pre-trained word2vec has 47 useful words (29. 44 %). The fine-tuned WangchanBERTa, the second most useful in the small-sized dataset, has the lowest number of useful words of 40 words (22.22 %). (χ2 test; $p < 0.05$).

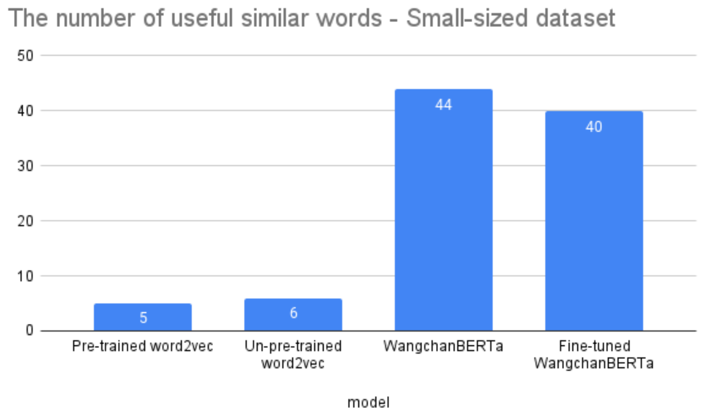


Figure 20: Number of useful similar words in delineating opposite opinion of each model of the small-sized datasets

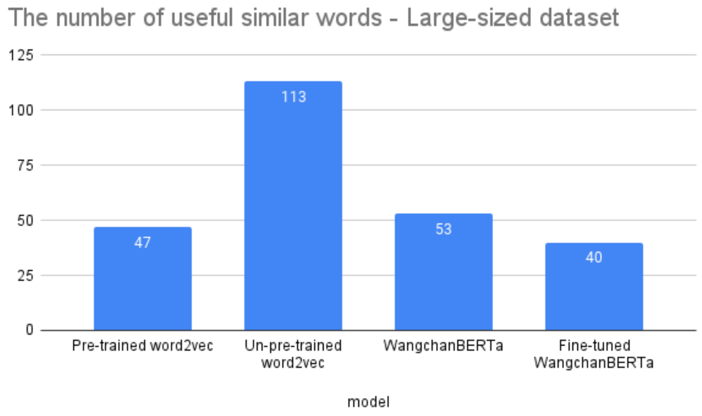


Figure 21: Number of useful similar words in delineating opposite opinion of each model of the large-sized datasets

The analysis above reaffirms the recommendation against utilizing word-based opinion analysis for small corpora due to the observed limitation in the number of useful words. Even the most effective model in the small corpus setting yields only 44 useful words, which is still lower than the count achieved by the least effective model in the large corpus, amounting to 47 useful words. Therefore, it is evident that this form of analysis is more suitable for a sufficiently large corpus, as it facilitates a more robust and reasonable opinion analysis.

The following is an example of the opposing opinion analysis utilizing un-pre-trained static word embedding on a large dataset. Regarding the protest, I can observe diverse opinions between the two groups. In terms of the demands of the protest - เรียกร้อง “demand/call for” (Table 3), both groups’ most similar words are the words right but differ in spelling; สิทธิ์ “right, e.g., a right to education” is of the protesters, and สิทธิ “right, e.g., a right to education” is of the opposition. However, the rests are different. The protesters' set focus on what they demand, such as พื้นฐาน “fundamental - sub word of fundamental rights”, สิ่ง “thing - usually used as the rightful thing in the corpus”, เป็นธรรม “fair - sub word of fairness”, ประชาธิปไตย “democracy”, ถูกต้อง “rightful”. On the contrary, the word เรียกร้อง “demand/call for” is similar to the word with negative meanings such as ละเมิด “infringe”, โจมตี “attack, assault”, ต่อต้าน “against”, and ยัดเยียด “force”. I can observe the text in the opposition corpus that say the protester called for their fundamental right, but they infringed on others' right.

All in all, contextualized word embedding is the only choice for a small corpus, while un-pre-trained word2vec is the greatest model for a large corpus.

3. How do static and contextualized word embeddings behave differently in word-based analysis?

In this section, I will investigate why the un-pre-trained static word embedding is more advantageous than the contextualized model and analyze the different behaviors of the static and contextualized word embeddings.

This section investigates the hypothesis concerning the association between the ability to distinguish between opposite corpora and the specific type of lexical relationship between a focus word and its similar words. To explore this hypothesis further, a categorization scheme is implemented, classifying each similar word within the corresponding set into three distinct categories: syntactic relation, semantic relation, and no relation. To exemplify this categorization, consider the focus word ม็อบ “protest”, which exhibits a syntactic relation with the similar word คาร์ “car” due to the modification of “protest” to form the phrase “car protest”. Conversely, the focus word รัฐบาล “government” and the similar word ฝ่ายค้าน “opposition” demonstrate a semantic relation as they both belong to the same semantic network. They are coordinated because they are under the same hypernym. Lastly, the focus word ประเทศ “country” has no discernible relation with the similar word ท้อแท้ “tired”.

Subsequently, the proportions of useful words within each specific relation category are computed. Figure 22 shows the percentage of useful words in each relation given the word relation. The analysis reveals that similar words exhibiting a syntactic relation exhibit the highest rate of usefulness, with 40.92% of these syntactically related words being deemed useful. In contrast, only 23.09% of the semantic-related words are classified as useful. Moreover, similar words lacking any discernible relation exhibit the lowest rate of usefulness, with only 13.19% of these words being considered valuable (χ2 test; $p < 0.05$). These findings underscore the favorable nature of syntactic relations in generating useful words for word-based opinion analysis, as evidenced in Figure 22.

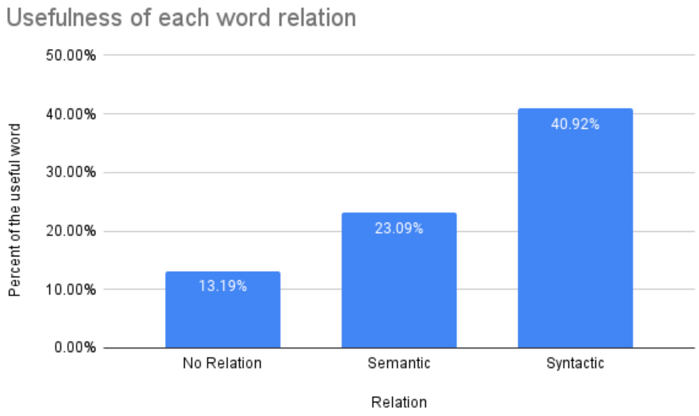


Figure 22: Usefulness of each word relation

I will begin with the semantic relation to demonstrate how each relation is useful in opinion analysis. There are some examples of useful similar words with semantic relations. The salient examples are those of the focus word ชาติ "nation". In both sets, there is a word แผ่นดิน "kingdom". This word also has a sense of father land, so both protesters and oppositions have the same opinion. From sampling text data, both groups have an emotional attachment to the word ชาติ "nation" because it relates to their father land. However, the differences show in the set of the opposition. The concept of ชาติ "nation" connects to the tradition and royal institution, expressed by the words วัฒนธรรม "culture", พระมหากษัตริย์ "king", and ศาสนา "religious". They also refer to the word ต่างชาติ "foreign/ foreigner", and ไทย "Thai", so their concept of a nation relates to the contrasting identity, themselves and others. On the contrary, the concept of ชาติ "nation" of the protesters is focused on within the boundary of Thai nationality and does not mention other nationality. In addition, some phrases express their anger.

There are similar words syntactically related to the focus word. For instance, the similar word ชั่ว “evil” usually forms a phrase with the focus word to be ชาติชั่ว “scumbag/bastard”. The similar word ปล้น “rob” is combined with ชาติ “nation” to be ปล้นชาติ “rob the nation”. This phrase shows the protesters feel the government is causing damage to the nation. In addition, the word กรรม “action/sin” are in the same syntactic tree, i.e., “to be born in the nation (ชาติ) is a sin (กรรม).” It expresses sorrows of the protesters. Another example is from the focus word รัฐบาล “government”, the similar word ส้นตีน “heel/damn” is formed with the focus word to be a phrase รัฐบาล ส้นตีน “damn government”. The similar word is a modifier of the focus word. The advantages of the syntactic relation are that it can observe some patterns or phrases showing the content in the text.

An example of a useful word that does not have any relation is ทหารเหี้ย “evil soldier”. The protesters expressed their anger due to the scandal associated with the Thai army. They feel that the evil government and soldiers cause the nation's decline.

The examples above illustrate how each relation is useful. However, the superior of the syntactic relation is not only because they show some phrases or patterns, but they usually represent the corpus. When sampling the text to analyze, it can be observed that the text in the corpus more align with the interpretation of the syntactic relation of similar words.

Examining lexical relationships within both the un-pre-trained and contextualized sets uncovered a distinct dissimilarity in the distribution of syntactic relationships. Notably, the un-pre-trained word2vec set demonstrated the highest proportion of syntactic relationships. In contrast, both WangchanBERTa and Fine-tuned WangchanBERTa exhibited the lowest proportion of syntactic relations, with a similar distribution pattern observed between the two (refer to Figure 23). This discrepancy suggests that the un-pre-trained model can differentiate between opposing corpora by generating similar words that exhibit syntactic relationships with the focus word. Consequently, these findings underscore the valuable insights the un-pre-trained model can provide for word-based opinion analysis tasks.

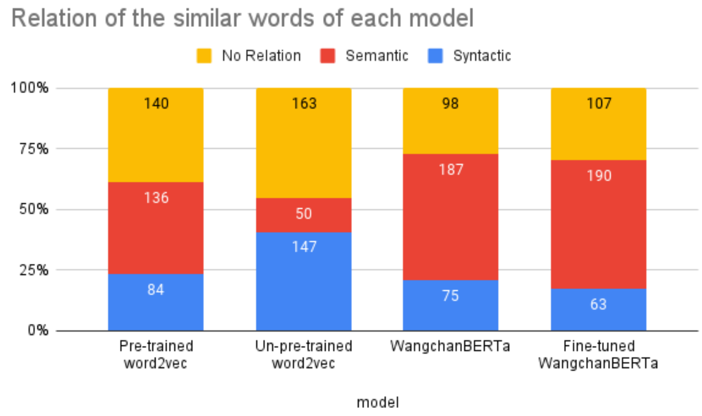


Figure 23: Relation of the similar words of each model

To do further analysis, I compare the top 2 most useful words embedding of each dataset. According to Figure 20, WangchanBERTa and Fine-tuned WangchanBERTa are the top 2 most useful models for small-sized dataset. Figure 24 shows that syntactic relation is the most useful word relation for delineating 2 corpora. (χ2 test; $p \geq 0.05$) The result of the large-sized dataset is aligned with the small dataset. (Figure 25) The un-pre-trained model's syntactic top relation makes it more beneficial for word-based opinion analysis than the contextualized model, where 60.6% of similar words have a semantic relationship expressing less corpus-specific information.

The contextualized model often yields the nearest neighbors with synonymous relation with the focus word, making it less suitable for opinion mining. This aligns with the findings of Bommasani et al. (2020), where contextualized embeddings performed best on synonym prediction. Unlike static embeddings, the contextualized model also tends to have nearest neighbors with the same part of speech as the focus word.

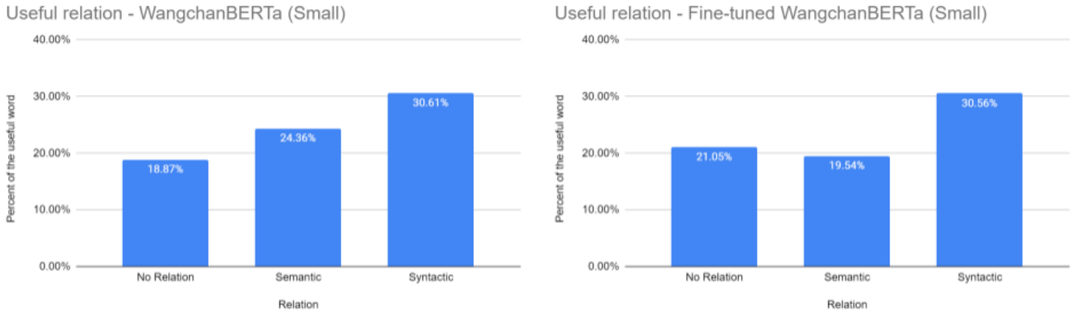


Figure 24: The comparison of useful relation of the top 2 useful word embedding of small-sized dataset

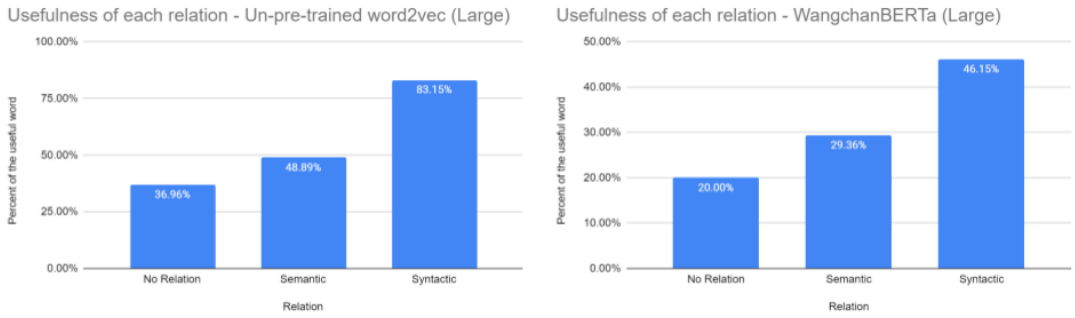


Figure 25: The comparison of useful relation of the top 2 useful word embedding of large-sized dataset

The following are examples to support this conclusion. Considering un-pre-trained sets, for example of the focus word ม็อบ "protest" (Table 7), the similar word คาร์ "car" modifies the focus word ม็อบ "protest". These words show that the protester conducted a "car mob". The similar word of the opposition corpus (Table 8), ป่วน "be turbulent", the focus word ม็อบ "protest" is an actor who makes the turmoil and turbulent. The focus word ม็อบ "protest" is a complement of the similar word, สลาย "disband". The opposition corpus contains texts about the crowd control police disbanding the protest. These syntactic relations give a clear picture of the opinion of each corpus.

On the contrary, the contextualized model gives similar words with semantic relations more than syntactic ones. There are several synonyms of the focus word ม็อบ "protest"; for instance, มวลชน "masses" (protester corpus), and การชุมนุม "assembly" (Table 8). Or the example of เรียกร้อง "demand/call for" from Table 3 and Table 4, there are troponyms, e.g., ทวง "demand, claim", entailment, e.g., ร้องขอ "ask for request, demand", and entailment, e.g., ถามหา "ask, inquire". These semantic words do not show the content in the focus corpus.

At this point, the major arguments hold. The static word embedding overfit the data, making it the best model for word-based analysis tasks. The contextualized word embedding pre-trained on the large dataset is less advantageous. It mostly gives the most similar word semantically related to the focus word.

At this point, the major arguments hold. The static word embedding overfit the data, making it the best model for word-based analysis tasks. The contextualized word

embedding pre-trained on the large dataset is less advantageous. It mostly gives the most similar word semantically related to the focus word.

4. To gain best practices for when and how to apply static and contextualized word embeddings in the analysis of two opposing opinions

The results confirmed that un-pre-trained word embedding is the most effective word-based opinion analysis for delineating opposing opinions. However, the un-pre-trained word embedding yields one embedding per word, so it cannot analyze polysemy. In addition, it requires the pre-determined separated corpora before training a separate word2vec model to create the word embeddings representing the corpora.

Therefore, this section aims to explore the feasibility of conducting word-based opinion analysis in scenarios where the text comprises a blend of opinions.

4.1 Clustering the instances of contextualized word embedding into 2 clusters to find whether the instances' embeddings can delineate opposite opinion corpora

First of all, contextualized word embedding yields one unique vector per instance of the word. For example, if the word "protest" occurs N times in total, N vectors represent the word protest. In this sub-section, I hypothesize that the word embedding of the words that appear in different corpora are significantly different. Thus, it is possible to cluster the embedding into 2 clusters. In addition, Figure 26 shows how to cluster each focus word's word embeddings into 2 clusters. There are N instances of a focus word. Each instance has its unique embedding. Then, all of them are fed into KMeans model. The results of the model are the cluster 0 or cluster 1 for each instance of focus word.

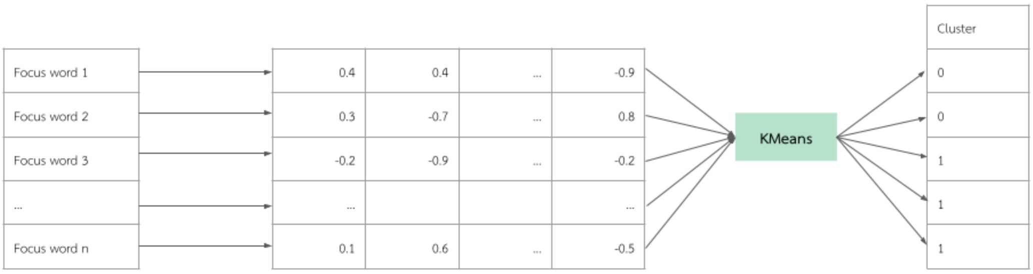


Figure 26: Process of clustering contextualized word embedding into 2 clusters

Following the clustering process, a reassignment of labels is conducted for the clusters identified by KMeans. Both possibilities are considered due to the potential ambiguity in interpreting cluster 0, representing either a protester or opposition corpus. In the initial evaluation, cluster 0 is labelled a protester corpus, while cluster 1 is designated an opposition corpus. Subsequently, the performance of the word embedding distinguishing between the opposing corpora is assessed using the macro-averaged F1 score as the evaluation metric. The actual corpus from which the embedding originates serves as the gold standard for this evaluation.

The obtained macro-averaged F1 scores for the first alternative are displayed in Table 9. The mean of macro-averaged F1 score for this alternative is 43.81, with a median value of 41.99 and a mode of 50.48. Notably, among the focus words, the third layer of the focus word ชาติ “nation” achieves the highest macro-averaged F1 score of 71.91, while the fourth layer of the focus word ประชาชน “people” attains the lowest score of 21.90.

The results of the alternative labelling are similar to the first labelling. The mean of macro-averaged F1 score is 40.97, the median at 41.03, and the mode at 30.05. The maximum score is the score of the eighth layer of the focus word ชาติ “nation”, which is 70.75, and the minimum score is the score of the fifth layer of the focus word ประชาชน “people”, which is 22.53.

In conclusion, it is impossible to differentiate the corpus from word embedding by clustering. Although the best performances have the macro-averaged F1 score over 70, it cannot conclude which layer is the best, and the rest layer performs poorly.

Focus word	1	2	3	4	5	6	7	8	9	10	11	12
แกง “deceive”	51.13	42.27	31.66	29.63	51.07	50.48	50.48	50.48	50.48	32.60	37.55	50.95
ชาติ “nation”	60.77	64.28	**71.91**	70.98	28.76	37.58	57.89	28.38	37.77	62.70	53.79	36.25
ประชาชน “people”	28.38	40.10	23.74	**21.90**	45.64	45.55	45.41	45.40	45.40	22.17	22.37	45.04
ประเทศ “country”	59.31	67.54	39.72	51.34	39.43	57.38	39.21	50.93	51.28	38.11	39.24	39.95
ภาษี “tax”	57.59	35.21	35.19	32.17	41.99	55.38	41.39	59.76	44.28	34.48	38.64	40.92
ม็อบ “protest”	40.50	39.56	25.65	46.76	48.91	40.01	48.79	40.71	40.93	40.95	43.47	44.48
รัฐบาล “government”	58.71	49.29	41.55	58.62	34.25	34.17	38.86	36.57	34.13	59.27	60.42	32.77
เรียกร้อง “demand/call for”	41.94	43.59	25.41	39.82	43.58	44.09	51.06	40.38	40.45	25.20	52.31	56.37
สลิ่ม “Sa-lim”	55.04	52.24	44.45	50.48	39.73	44.37	39.67	39.81	39.61	43.34	45.03	40.46

Table 9: macro-averaged F1 score of the labelling scheme: 0 as Protester and 1 as Opposition

Focus word	1	2	3	4	5	6	7	8	9	10	11	12
แกง “deceive”	35.41	43.75	49.77	50.78	29.20	30.05	30.05	30.05	30.05	49.46	46.12	30.00
ชาติ “nation”	29.22	27.72	26.76	26.68	64.96	55.95	41.45	**70.75**	56.17	30.17	42.23	47.02
ประชาชน “people”	50.13	36.73	44.92	45.61	**22.53**	23.19	23.90	23.94	23.94	46.04	45.94	25.29
ประเทศ “country”	33.76	29.49	39.86	37.68	39.34	40.24	40.40	46.41	46.05	39.36	56.99	41.67
ภาษี “tax”	31.92	51.72	43.13	58.06	37.56	34.23	38.78	30.45	32.75	43.47	41.45	40.24
ม็อบ “protest”	40.76	42.85	48.73	38.22	25.17	43.64	25.44	41.72	41.05	41.03	39.81	39.42
รัฐบาล “government”	35.24	40.49	38.37	34.90	59.72	59.87	41.01	55.67	59.88	34.43	33.81	62.02
เรียกร้อง “demand/call for”	44.11	47.67	50.96	46.99	44.63	42.07	25.66	50.09	48.48	50.36	38.55	34.61
สลิ่ม “Sa-lim”	42.38	43.10	40.61	42.87	44.03	50.42	44.71	44.72	44.15	38.88	39.68	44.64

Table 10: macro-averaged F1 score of the labelling scheme: 0 as Opposition and 1 as Protester

4.2 Clustering the instances of contextualized word embedding into multiple clusters to find whether it is possible to study polysemy from word embedding

The clustering process follows the methodology depicted in Figure 26, where the number of clusters is determined using the elbow method based on the distortion metric. Subsequently, the KMeans model assigns cluster labels to each instance of the focus word. A comparison is then made between the instance word embedding and the average contextualized word embedding, allowing for identifying the most similar words (as illustrated in Figure 27). These similar words' frequencies within the cluster are subsequently recorded. Figure 28 is an illustrative example, showcasing the process of tallying the frequency of similar words within Cluster 0 of a given focus word. The top 10 most frequently occurring similar words represent the cluster, contributing to interpreting the cluster's meaning.

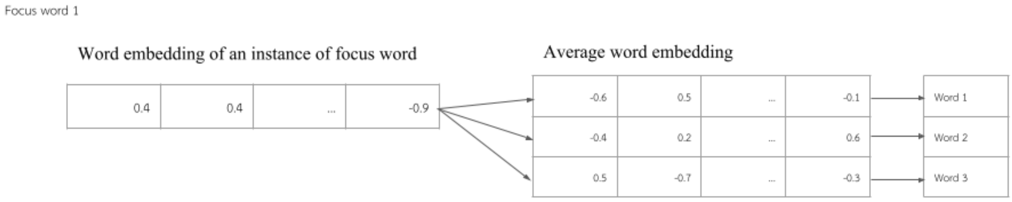
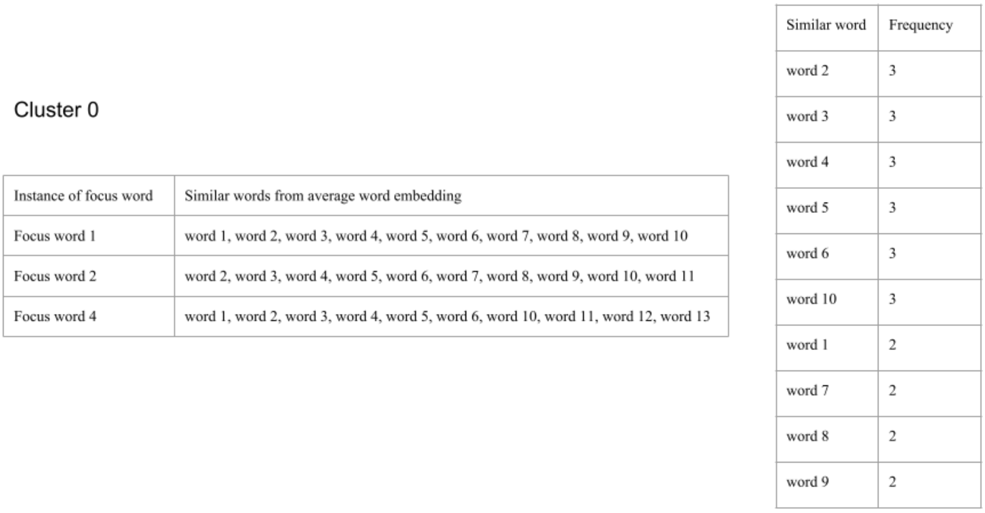


Figure 27: Process of finding the most similar words for each instance of a focus word



Cluster 0

Instance of focus word	Similar words from average word embedding
Focus word 1	word 1, word 2, word 3, word 4, word 5, word 6, word 7, word 8, word 9, word 10
Focus word 2	word 2, word 3, word 4, word 5, word 6, word 7, word 8, word 9, word 10, word 11
Focus word 4	word 1, word 2, word 3, word 4, word 5, word 6, word 10, word 11, word 12, word 13

Similar word	Frequency
word 2	3
word 3	3
word 4	3
word 5	3
word 6	3
word 10	3
word 1	2
word 7	2
word 8	2
word 9	2

Figure 28: Example of how to count the frequency of similar words in Cluster 0 of a focus word

The interpretations of clusters for each focus word are in Table 11.

Focus word	Cluster	Interpretation
แกง “deceive”	0	Meaning related to the protest
	1	Meaning related to the protest
	2	Meaning related to the protest
	3	Meaning related to food
	4	Meaning related to food
	5	Not related to both food and the protest
	6	Not related to both food and the protest
ชาติ “nation”	0	Meaning related to the institution
	1	Abuse words
	2	Meaning related to life, e.g., ชาตินี้ “this life”, ชาติหน้า “next life”, ชาติที่แล้ว “past life”
	3	Race
	4	National team
ประชาชน “people”	0	Meaning related to people
	1	Meaning related to people
	2	Meaning related to people
	3	Meaning related to people
	4	Meaning not related to the focus word
	5	Meaning not related to the focus word
ประเทศ “country”	0	Country, state, nation
	1	Administrative district
	2	Country, state, nation
	3	Country + conj.
	4	Country + conj.
ภาษี “tax”	0	Right from paying tax
	1	Tax fraud
	2	Government budget, VAT
	3	Bribe
ม็อบ “protest”	0	Name of the protest event, making an appointment
	1	Arrest, confrontation
	2	Protester, mainstay
	3	Alliance
	4	Meaning not related to the protest
รัฐบาล “government”	0	Government
	1	Government
	2	Government
	3	Meaning not related the focus word
	4	Meaning not related the focus word
เรียกร้อง “demand/call for”	0	Protest, fight for
	1	Protest, fight for
	2	Protest, fight for
	3	Protest, fight for
	4	Protest, fight for
	5	Protest, fight for
	6	Meaning not related to the focus word
สลิ่ม “Sa-lim”	0	Meaning related to the focus word
	1	Meaning related to the focus word
	2	Meaning related to the focus word
	3	Meaning related to the focus word
	4	Meaning related to the focus word
	5	Meaning related to the focus word
	6	Meaning related to the focus word

Table 11: The interpretation of the cluster

According to the Table 11, ชาติ “nation”, ภาษี “tax”, and ม็อบ “protest” are the focus word that shows polysemy distinctly. Some of the focus words have few senses than the number of clusters, e.g., แกง “deceive”, ประชาชน “people”, ประเทศ “country”, รัฐบาล “government”, and เรียกร้อง “demand/call for”. สลิ่ม “Sa-lim” is the only word that cannot detect polysemy from the interpretation.

Subsequently, the weighted proportion of corpora within each cluster is calculated. The corresponding charts illustrate these proportions for all focus words in the appendix. The observed distribution of corpora within the clusters reveals that distinct corpora exhibit varying distributions of subsenses. This discrepancy in distribution suggests the presence of divergent opinions within the opposing corpora. Thus, it can be inferred that the dissimilar distributions of subsenses serve as evidence of differing opinions among the opposite corpora.

A text sampling process is employed for each cluster to ensure the accuracy of interpretation. The findings reveal that for 8 out of 9 focus words, there are no discernible differences in meanings between the clusters. Consequently, the interpretation can only be verified for a single focus word, namely ชาติ “nation”. Most of the sampled texts within this cluster align with the proposed interpretation. However, it is noteworthy that the meaning conveyed by the sequences of words differs substantially between the two corpora. For instance, in cluster 1, the usage of the word ชาติ “nation” is associated with abusive language. Both corpora employ the term ชาติชั่ว “scumbag/bastard” as an abusive expression, yet the specific topics surrounding these derogatory terms diverge between the opposing corpora.

Regarding the 8 out of 9 focus words that do not have different meanings, the possible explanation for this situation is that the word embedding encodes a slight difference in meaning, although the texts are not different.

In conclusion, contextualized word embedding can depict some polysemy but should not be used for delineating opposing corpus.

4.3 Using the sequence embedding from WangchanBERTa to cluster opposite opinions and studying opinions from the word2vec model

Based on the findings presented in Section 4.2, it was observed that the disparity between corpora primarily manifests at the sequence level rather than the word level. Consequently, clustering was performed using sequence embeddings to segregate the distinct opinions. WangchanBERTa provides the CLS token embedding, which effectively captures the overall meaning of the sequence. Furthermore, as indicated by the outcomes discussed in Section 2, the un-pre-trained word2vec model emerged as the most suitable model for distinguishing opposing opinions. Following the clustering of sequence embeddings into two distinct clusters, a word2vec model was trained on each text cluster. Subsequently, a comparison was conducted between the most similar words derived from the newly trained word2vec model and the un-pre-trained word2vec model utilized for the original corpus analysis.

The clustering methodology employed in this study remains consistent with the approach described in the preceding section. Specifically, the KMeans algorithm was utilized with a designated number of clusters set to two. The subsequent step involved relabeling the clusters in two distinct ways: assigning cluster 0 to represent the protester corpus and cluster 0 to represent the opposition corpus. The effectiveness of the word embedding in distinguishing between these opposing corpora was subsequently evaluated using the macro-average score as a performance metric. The evaluation process involved comparing the embeddings with the respective actual corpus from which they originated, thereby utilizing the actual corpus as a gold standard for assessment.

The evaluations of both labelling show that clustering sequence embedding performs poorly. The first labelling scheme has the macro-averaged F1 score of 36.76, and that of the second labelling is 33.67, lower than 50. Thus, I opt not to train word2vec from the clustered text as the clustering is inaccurate.

4.4 Using the WangchanBERTa classifier to predict the corpus of text and training word2vec from the predicted corpus

As Figure 5 (b) describes, sequence classification is a downstream task of WangchanBERTa. I expect the classification to yield more accurate results than the clustering in 4.3.

This experiment is done in the following steps.

1. Sampling 4,000 tweets from each corpus and splitting again into 3,000, 500, and 500 for train, validation, and test sets.
2. Combining the train set of both corpora, repeat this step on the validation and test sets.
3. Shuffling the tweet in all three sets.
4. Training the classifier model to predict the corpus and evaluate the classifier on the test set.
5. If the macro average of the prediction result of the test set is over 80, this classifier is used to predict the rest of the text (56,292)
6. Evaluating the predicted set again to ensure the macro average is above 80.
7. If the macro average is above 80, combine the text with the same corpus label.
8. Training word2vec on each corpus with the same parameter as un-pre-trained word2vec.

The classifier used in this study was trained using the WangchanBERTa model. The input to the classifier consisted of the CLS (classification) token. The training process involved training the classifier for a total of four epochs.

Overall, the classifier considerably outperforms the clustering in 4.3. The macro-averaged F1 score of the test set is 83. The evaluation of the rest of the text (56,292) is 84. Therefore, this prediction should be accurate enough.

	precision	recall	f1-score	support
Protester	0.8187	0.9005	0.8576	30,186
Opposition	0.8699	0.7694	0.8166	26,106
accuracy	0.8397	0.8397	0.8397	0.8397
macro avg	0.8443	0.8349	0.8371	56,292
weighted avg	0.8424	0.8397	0.8386	56,292

Table 12 : The report for classifier evaluation on the rest of the text (56,292)

Subsequently, the word2vec model was trained using the predicted corpora, and the most similar words to the focus words were obtained. These sets of similar words from the predicted corpus were then compared with the sets of similar words generated by the un-pre-trained word2vec model trained on the original corpus. In this comparison, the un-pre-trained word2vec model of the original corpus served as the gold standard, as it was established in section 2 that this model outperformed others for word-based opinion analysis.

The correctness score was computed to ensure an unbiased evaluation. Each similar word was assessed, and if it was found to be present in the gold standard set, the score was incremented by 1. Subsequently, the total score was divided by 10. The correctness scores obtained are presented in Table 13.

Focus word	Protester	Opposition
แกง “deceive”	0.6	0.3
ชาติ “nation”	0.6	0.8
ประชาชน “people”	0.5	0.4
ประเทศ “country”	0.8	0.9
ภาษี “tax”	0.7	0.9
ม็อบ “protest”	0.9	0.2
รัฐบาล “government”	0.6	0.5
เรียกร้อง “demand/call for”	0.6	0.3
สลิ่ม “Sa-lim”	0.5	0.2
Average	0.64	0.5

Table 13: The correctness score of the similar word sets from the predicted corpus

In general, the word2vec predictions associated with the protester exhibit a higher degree of similarity to the established gold standard than those associated with the opposition. Specifically, all sets of focus words related to the protester consistently attain a score of at least 0.5, whereas the opposition's minimum correctness score is 0.2. Consequently, the protester demonstrates an average correctness score of 0.64, while the opposition achieves an average of 0.5.

Thought the correctness score is not that impressive, the similar words from word2vec shown in Table 14 and Table 15 show that the predicted corpus is still useful. The meanings of words in similar sets of both gold standard word2vec and predicted word2vec are not diverse. For example, the focus word เรียกร้อง "demand/call for" of protester corpus has the correctness score of 0.6, but it is still possible to recognize that the text in the predicted corpus is about demanding fairness. The other example is the focus word ม็อบ "protest" from the opposition corpus. This focus word correctness score is only 0.2, but it is still possible to observe a negative opinion toward the protest from the opposition by the words บุก "trespass", and กุ๊ย "tramp". It seems that the opposition think that the protesters are the cause of the unrest.

Focus word	Word2vec (Gold standard)	Word2vec from predicted corpus
แกง "deceive"	หม้อ "pot", ยูธ "subword of the group's name, free youth", แผน "plan", แกนนำ "leader, mainstay", แก "you", โอเค "OK", รับผิดชอบ "be responsible", เตือน "remind, warn", สถานที่ "location", จัด "arrange"	หม้อ "pot", ยูธ "subword of the group's name, free youth", โอเค "OK", แก "you", สรุป "conclude", รับผิดชอบ "be responsible", คุย "talk", สถานที่ "location", ฟรี "free", หลอก "deceive"
ชาติ "nation"	ชัง "hate", ชั่ว "evil", ทหารเหี้ย "bad soldier", แผ่นดิน "kingdom", ม็อบสามกีบ "three-toed hoof protest", ปรสิต "parasite", แท้ "authentic", อุบาทว์ "wicked", ตระกูล "lineage", ปล้น "rob"	ชั่ว "evil", ชัง "hate", ม็อบสามกีบ "three-toed hoof protest", แผ่นดิน "kingdom", ศาสนา "religious", ทหารเหี้ย "bad soldier", ตระกูล "lineage", ตกนรก "go to hell", ประชา "people", เปรต "hungry ghost"
ประชาชน "people"	ทำร้าย "harm", ประชาชนอะ "people", ปชช. "Abbreviation of people", ตอบแทน "repay", ประชาชนอ่ะ "people", ปปช. "anti-graft panel", ศัตรู "enemy", ใช่ "yes", ปกป้อง "protect", หรอ? "Really?"	ทำร้าย "harm", ประชาชน! "people", ตอบแทน "repay", พระราชา "king", ปชช. "Abbreviation of people", รังแก "bully", ประชาชนอ่ะ "people", เนรคุณ "betray", ประชาชนอะ "people", ชั่วช้า "evil"

Focus word	Word2vec (Gold standard)	Word2vec from predicted corpus
ประเทศ “country”	ตกต่ำ “decline”, ประเทศชาติ “nation”, เจริญ “prosper”, ประเทศเหี้ย “damn country”, เน่าเฟะ “putrid”, ปท “country”, ถ่วง “hinder”, ท้อแท้ “tired”, ซวย “unfortunate”, พัฒนา “develop”	พัฒนา “develop”, เจริญ “prosper”, เน่าเฟะ “putrid”, ประเทศเหี้ย “damn country”, ซวย “unfortunate”, ประเทศชาติ “nation”, ถ่วง “hinder”, ดักดาน “unmitigated”, ประเทศเฮงซวย “damn country”, ตกต่ำ “decline”
ภาษี “tax”	จ่าย “pay”, เดือน “month”, เงิน “money”, ตอบแทน “repay”, แดก “devour”, แดก “devour”, ซื้อ “buy”, สวัสดิการ “welfare”, เนรคุณ “betray”, ตัง “money”	จ่าย “pay”, เงิน “money”, เดือน “month”, จ้าง “employ”, ตอบแทน “repay”, เปลือง “waste”, แดก “devour”, เนรคุณ “betray”, แดก “devour”, พระราชา “king”
ม็อบ “protest”	ธรรมเนียบรัฐบาล “Government House”, วัน “day”, ,rt , Please, , สาธารณะสุข “public health”, Retweet, หาดใหญ่ “Hat Yai”, ธรรมเนียบ “residence”, 💢	สาธารณะสุข “public health”, ธรรมเนียบรัฐบาล “Government House”, Please, ,rt, Retweet, 💢, หาดใหญ่ “Hat Yai”, กระจ่าย (misspelled word), , วัน “day”
รัฐบาล “government”	เฮงซวย “bad”, ส้น “heel”, ตีน “foot”, ล้มเหลว “fail”, ม็อบส้น “damn protest”, ประเทศส้น “damn country”, รบ. “Abbreviation of government”, ส้นตีน “heel/damn”, เชียร์ “cheer”, อีรัฐบาล “damn government”	รบ. “Abbreviation of government”, เฮงซวย “bad”, ประเทศส้น “damn country”, ส้น “heel”, รัด “sub-word of misspelled of government/bind”, ตีน “foot”, ส้นตีน “heel/damn”, อีร้า “sub-word of misspelled of government”, งาม “be beautiful”, ผิดพลาด “be erroneous”
เรียกร้อง “demand/call for”	สิทธิ์ “right”, พื้นฐาน “fundamental”, แค่ “only”, สิ่ง “thing”, เป็นธรรม “be fair”, ประชาธิปไตย “democracy”, ถูกต้อง “right”, ด้วยซ้ำ “as well”, รับฟัง “listen to”, ผิดหรอ “Is it wrong?”	พื้นฐาน “fundamental”, สิทธิ์ “right”, เสรีภาพ “freedom”, เท่าเทียม “be equal”, เป็นธรรม “be fair”, จุดยืน “standpoint”, ผิดหรอ “Is it wrong?”, เขา “he/she/they”, ถูกต้อง “right”, ด้วยซ้ำ “as well”
สลิ่ม “Sa-lim”	เบิกเนตร “open eyes”, กะลา “coconut shell”, แหกตา “Wipe your eyes.”, บูด “rancid - it can be used as slang meaning bad”, โทรทัศน์ “TV”, มัว “be engrossed in/be busy with”, ลืมหูลืม “open eyes”, หลับหูหลับ “to be blind (in belief)”, ก้ “also”, รำคาญ “feel annoyed”	แหกตา “Wipe your eyes.”, กะลา “coconut shell”, เฟส “face (may be from Facebook)”, ลืมหูลืม “open eyes”, เบิกเนตร “open eyes”, มัว “be engrossed in/be busy with”, กี “misspelled pussy”, ทีเหอะ “phrase that cannot translate without contex”, ดิ้น “wriggle”, ดักดาน “unmitigated”

Table 14: The similar word sets of gold standard protester and predicted protester corpus

Focus word	Word2vec (Gold standard)	Word2vec from predicted corpus
แกง "deceive"	เจียว "render", หอม "shallot", ทอด "fry", หม้อ "pot", 😋, หั่น "cut", ปลา "fish", ยยยย "a cluster of a character", นม "breast, milk", ไก่ "chicken"	ยยยย "a cluster of a character", กุ้ง "shrimp", ราด "pour", กล้วย "banana", ขน "carry", ชาม "bowl", เจียว "render", ปา "throw", เผ็ด "spicy", หั่น "cut"
ชาติ "nation"	ชัง "hate", แผ่นดิน "kingdom", ไทย "Thai", ต่าง "different", ธง "flag", วัฒนธรรม "culture", พระมหากษัตริย์ "king", กษัตริย์ "king", ศาสนา "religious", ประเทศ "country"	ชัง "hate", ธง "flag", พระมหากษัตริย์ "king", แผ่นดิน "kingdom", ไทย "Thai", ศาสนา "religious", เพศ "gender", ศิลปิน "artist", กษัตริย์ "king", วัฒนธรรม "culture"
ประชาชน "people"	จัดการ "manage", เด็ดขาด "be decisive", ปชช "Abbreviation of people", กฎ "rule", เพื่อ "for", สมควร "suitably", เป้าหมาย "goal", กระบวนการ "process", ล้มเหลว "fail", อัยการศึก "martial law"	นโยบาย "policy", ปชช "Abbreviation of people", ปชช. "Abbreviation of people", บริหารจัดการ "manage", ล่ม "fail", ทั่วถึง "throughout", เป้าหมาย "goal", กระบวนการ "process", ทรัพยากร "resource", จัดการ "manage"
ประเทศ "country"	พัฒนา "develop", ต่าง "different", เมืองไทย "Thailand", ไทย "Thai", ปท "country", ร่ำรวย "rich", มหาอำนาจ "powerful", ประเทศไทย "Thailand", โลก "world", เพื่อนบ้าน "neighbor"	พัฒนา "develop", ทั่ว "throughout", ร่ำรวย "rich", มหาอำนาจ "powerful", เพื่อนบ้าน "neighbor", ต่าง "different", ปท "country", ไทย "Thai", เมืองไทย "Thailand", ประเทศไทย "Thailand"
ภาษี "tax"	จ่าย "pay", เงิน "money", หนี้ "debt", จ้าง "employ", เยียวยา "treat", แพง "expensive", สวัสดิการ "welfare", คุ้ม "be worthwhile", ประกัน "guarantee", กำไร "profit"	สวัสดิการ "welfare", จ่าย "pay", เงิน "money", เยียวยา "treat", หนี้ "debt", กำไร "profit", คุ้ม "be worthwhile", จ้าง "employ", ลดหย่อน "(tax) credit", แพง "expensive"
ม็อบ "protest"	ม๊อบ "protest", ชุมนุม "gather", ป่วน "be turbulent", สลาย "disband", ตำรวจ "police", เผา "burn", ฝูง "herd", ดินแดง "Din-Dang", คฝ "Crowds Control polices", แกนนำ "leader, mainstay"	ม๊อบ "protest", ขีด "mark", สลาย "disband", บุก "trespass", ประณาม "condemn", อภิสิทธิ์ "Abhisit", ร่าน "crave", แถ "to beat around the bush", ถล่ม "bombard", กุ๊ย "tramp"
รัฐบาล "government"	รบ. "Abbreviation of government", รบ "Abbreviation of government", ห่วย "be inferior", จัดการ "manage", ประชาชน "people", ฝ่าย "side", สส "representative", เฮงซวย "bad", โจมตี "attack, assault", เด็ดขาด "be decisive"	รบ. "Abbreviation of government", ห่วย "be inferior", บริหาร "manage", เฮงซวย "bad", รัด "sub-word of misspelled of government/bind", ด่าว่า "scold", ประยุทธ์ "Prayut", โจมตี "attack, assault", รบ "Abbreviation of government", ล้มเหลว "fail"

Focus word	Word2vec (Gold standard)	Word2vec from predicted corpus
เรียกร้อง “demand/call for”	สิทธิ “right”, อ้าง “refer to, claim”, ละเมิด “infringe”, กระบวนการ “process”, โจมตี “attack, assault”, เสรีภาพ “freedom”, ต่อต้าน “resist”, มนุษยชน “human”, ยัดเยียด “force”, กฎหมาย “law”	มนุษยชน “human”, เท่าเทียม “be equal”, เป้าหมาย “goal”, ตำหนิ “flaw”, ต่อต้าน “resist”, ปกป้อง “protect”, ตามใจ “indulge”, อ้าง “refer to, claim”, สมเหตุสมผล “reasonably”, เชื่อมั่น “trust in”
สลิ่ม “Sa-lim”	ทัวร์ “tour”, อวย “praise”, ด้อม “fandom”, กรู “I”, อาย “be ashamed”, แท็ก “tag”, ผี “ghost”, โควท “quote”, งง “confuse”, เถียง “argue”	สามกีบ ““three-toed hoof”, โพส “post”, เถียง “argue”, ฉาม “misspelled three”, lll, เม้น “comment”, บล็อก “block”, ตอแหล “lie”, อิพวก “idiot”, กรู “I”

Table 15: The similar word sets of gold standard opposition and predicted opposition corpus

In conclusion, the optimal approach for employing contextualized word embedding and static word embedding in opinion analysis is to leverage a combination of both models. Contextualized word embedding offers the advantage of minimizing manual efforts, particularly when dealing with unclassified corpora. This type of model exhibits the ability to classify text into distinct categories. Conversely, static word embedding, such as word2vec, excels in extracting opinion based on individual words by examining their similar counterparts. By employing a hybrid approach, researchers can benefit from the strengths of each model, thereby enhancing the overall effectiveness of opinion analysis.

5. Example of opinion analysis utilizing un-pre-trained static word embedding

In the last section of this study, I will give an example of opinion analysis by utilizing the best word embedding model, which is un-pre-trained static word embedding. The analysis will follow the topics of interest stated in the result section's beginning.

The first topic is the protest itself, which consists of 3 focus words, i.e., แกง “deceive”, เรียกร้อง “demand/call for”, and ม็อบ “protest”. For the protesters, they used the Twitter platform to organize the demonstration. For instance, the focus word ม็อบ “protest” set contains similar words of location and date of the protest, such as ธรรมเนียบรัฐบาล “Government House”, สาธารณะสุข “public health”, หาดใหญ่ “Hat Yai”,

ธรรมเนียบ “residence (shorten for the government house)”, and วัน “day”. Twitter is the main platform to promote the event, so some words show the characteristic of the platform, e.g., Please, “,rt”, Retweet, and emoji 💢. There are a huge proportion of the texts, “Please, retweet วันนี้มีม็อบ” (Please retweet, there is a protest today).

The focus word แกง “deceive” have similar words แผน “plan”, แกนนำ “leader, mainstay”, สถานที่ “location”, and จัด “arrange”. From the interpretation and text sampling, protesters talked about the leaders tricking the คฝ “Crowds Control polices” to go to the wrong location. Some also criticized that the leaders must be responsible for the tricked and arrested protesters.

For the last word เรียกร้อง “demand/call for”, it can be observed that the protesters demand the fundamental right and fairness from the similar words สิทธิ์ “right”, พื้นฐาน “fundamental”, เป็นธรรม “be fair”, and ถูกต้อง “right”. Some similar words can be combined to form a phrase, i.e., สิ่ง “thing” and ถูกต้อง “right” are in the phrase สิ่งที่ถูกต้อง “rightful thing”. They also demand ประชาธิปไตย “democracy”.

On the contrary, the opposition has a negative opinion of the protest. According to the set of ม็อบ “protest”, the set contains ป่วน “be turbulent”. The similar word ฝูง “herd” convey the negative perspective since it shows that the oppositions see the protesters were uncivilized. They mention the unrest in the event, such as เผา “burn”, and สลาย “disband”. The opposition said that the police and Crowds Control police should disband the protest due to its disorder.

For the focus word เรียกร้อง “demand/call for”, there is an overlapping word, สิทธิ “right”, but the rest show conflicts. The set contains ละเมิด “infringe”, ต่อต้าน “resist/against”, and ยัดเยียด “force”. The overall interpretation is that the protesters call for their rights, but it infringes on others.

For the focus word แกง “deceive”, the set contains general meaning related words, e.g., เจียว “render”, หอม “shallot”, ทอด “fry”, and หม้อ “pot”. Though some of the text contains the sense of “deceive”, it is not a big proportion. Overall, the opposition does not mention the tactic of the protest.

In Conclusion, for the topic of the protest, the protesters focused mainly on the organization of the demonstration and their demands, while the opposition expressed dissatisfaction due to the unrest.

For the second topic, the protesters' opinion toward the opposition, the protesters see the opposition as those living in a shell. There are words กะลา “coconut shell”, มัว “be engrossed in/be busy with”, หลับหูหลับ(ตา) “to be blind (in belief)”, and โทรทัศน์ “TV”. There are texts, “Sa-lim should come out from the coconut shell”. The word โทรทัศน์ “TV” is the main source of information for the opposition. The protesters also criticized the government for censoring the news in mainstream media. Thus, if people consume news only through TV, they will have limited eyesight and not empathize with others. That is why “Sa-lim” oppose the protest.

On the other hand, the opposition joke about the critics. The opposition admits that they flood the negative comments on some tweets and make fun that the protesters also flood the negative comments on their tweets. They also admit that they praise the government because the government does their best. These interpretations are from the similar words ทัวร์ “tour”, and อวย “praise”. The word ทัวร์ “tour” is from the phrase ทัวร์ลง “backlash”. These are initially uninterpretable, but I can use them to sample text to analyze.

Overall, both groups seem to have severely opposite opinions.

The third topic is the concept of nation, ชาติ “nation”. Both groups have an overlapping word, e.g., ชัง “hate”, and แผ่นดิน “kingdom/homeland”. The word ชัง “hate” is from the phrase ชังชาติ “the haters of the nations”. This phrase is first used by the opposition to call the protesters since the protesters mostly criticize the government’s management and the royal family, which represent the values of the nation for the opposition. The protesters feel unfair. Both have similar opinions that the nation is related to their homeland.

Nevertheless, there are huge differences. The focus word ชาติ “nation” of protesters has negative similar words illustrating the nation is in turmoil, expressed by the word ปล้น “rob”. It is from the phrase ปล้นชาติ “rob the nation”. Some words refer to the entity that does damage to the nation, e.g., ทหารเหี้ย “bad soldier”, ปรสิต “parasite”, and ตระกูล “lineage”. They also use the focus word to express their anger since ชาติ “nation” also means “life” and is part of an abusive word such as ชาติชั่ว “scumbag/bastard”.

On the contrary, the idea surrounding the concept of a nation of the opposition is the tradition, religious, and royal family, which can be observed from the similar words วัฒนธรรม “culture”, พระมหากษัตริย์ “king”, กษัตริย์ “monarch”, and ศาสนา “religious”. Their values of the nation come from the ideology, ชาติ ศาสนา พระมหากษัตริย์ “nation, religious, king”. Therefore, the critics of one of these entities will irritate their feeling.

Comparing the sets of the focus word ประเทศ “country”, I found contrasting opinions—the protester's despair of the country. The protesters’ set consists of ตกต่ำ “decline”, ประเทศเหี้ย “damn country”, เน่าเฟะ “putrid”, ถ่วง “hinder”, and ท้อแท้ “tired”. Though there are some positive words, such as เจริญ “prosper” and พัฒนา “develop”, they used to express their hopes. In contrast, the opposition focuses on different aspects. They focus on the pros of Thailand compared to foreign countries. In a similar word set, the opinion cannot be observed directly. The words เมืองไทย “Thailand”, ไทย “Thai”, and ประเทศไทย “Thailand” are the keywords to search, and sampling text to analyze. They usually compare Thailand with a prosperous country. (พัฒนา “develop”, ร่ำรวย “rich”, มหาอำนาจ “powerful”)

The set of focus word ประชาชน “people” of the protesters presents the despair. The set contains ทำร้าย “harm”, and ศัตรู “enemy”. From the test, protesters said that the government acts as the enemy of the people and war against people. This is due to the crowd control police attack protesters. Conversely, the opposition agrees to the action of the police because it is legitimate to do so, i.e., จัดการ “manage”, สมควร “suitably”, กฎ “rule”, and อัยการศึก “martial law”.

Lastly, the topic of political policy and entity contains ภาษี “tax” and รัฐบาล “government”. For protesters it is obvious that the protesters feel that they are betrayed according to the similar word เนรคุณ “betray.” Their tax is a salary of the government. Still, the government enjoys spending their tax, i.e., จ่าย “pay”, เงิน “money”, เดือน “month” are combined to form a phrase จ่ายเงินเดือน “pay your salary”. The similar words แดก “devour” and แดก “devour” mean that the government greedily consumes the tax instead of use on welfare.

On the other hand, the opposition set also has the word จ่าย “pay”, เงิน “money”, สวัสดิการ “welfare”. However, when considering the sample text, their focus is on the welfare requires expensive tax, but Thailand is still indebted, and people pay less tax, such as หนี้ “debt”, แพง “expensive”.

Regarding the focus word รัฐบาล “government”, the protesters have various negative similar words, such as เฮงซวย “bad”, ส้น “heel”, ตีน “foot”, ล้มเหลว “fail”, ม็อบส้น “damn protest”, ประเทศส้น “damn country”, ส้นตีน “heel/damn”, and อีรัฐบาล “damn government”. They express their anger towards the government and the people who still support the evil government. On the other hand, the opposition also has a negative word, but the salient word is เด็ดขาด “be decisive”. The opposition is not satisfied that the government did not disband the protest absolutely.

In conclusion, some differences between opinions can be observed by analyzing the similar word set. According to the analysis, both groups have opposing opinions on every topic. There seems to be a long way to go for understanding between both groups.

CONCLUSION

Within the scope of this investigation, the research questions about word-based opinion analysis are addressed. Firstly, it is observed that pre-training significantly influences the quality of embeddings in datasets with limited size. Sufficient training data is crucial for shifting random weight embeddings toward higher-quality representations. Improved quality is evident in similar words that connect meaning to the focus word. However, conducting opinion analysis tasks on small-sized datasets is not recommended, as the limited data fails to facilitate the desired weight shift necessary for extracting corpus-specific information. Conversely, pre-training does not enhance embedding quality in large-sized datasets. This can be attributed to the vast amount of pre-training data overwhelming the corpus-specific data, resulting in noise rather than improvement when additional training is conducted.

Secondly, word-based opinion analysis is not advisable for small-sized datasets. Although contextualized word embedding (WangchanBERTa) yields more useful words than other models, this count remains below that of the worst-performing model in the large dataset. Consequently, word-based opinion analysis is recommended for large-sized datasets, with un-pre-trained static word embedding emerging as the most superior model.

Thirdly, static word embedding (Word2vec) is preferable over contextualized word embedding (WangchanBERTa) due to its ability to provide syntactically related similar words that aid in distinguishing opposing opinion corpora. Word2vec offers similar words syntactically linked to the focus word, while WangchanBERTa provides semantically related similar words.

Lastly, the most effective approach for utilizing static and contextualized word embeddings in analyzing two opposing opinions involves employing contextualized sequence embedding to predict the text corpus, followed by training a Word2vec model using the predicted corpus. Subsequently, the corpus analysis is performed based on the most similar words generated by the Word2vec model.

REFERENCES

Aroonmanakun, W., Tansiri, K., & Nittayanuparp, P. (2009, 2009/08/06/). Thai National Corpus: a progress report.*ALR7*

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, *131*(4), 1593-1636. https://doi.org/10.1093/qje/qjw024

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*. http://arxiv.org/abs/1607.04606

https://arxiv.org/abs/1607.04606

Bommasani, R., Davis, K., & Cardie, C. (2020, 2020/07//). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. ACL 2020,

Cat. (2023). In *Wikipedia*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805v2

https://arxiv.org/abs/1810.04805

Di Porto, F., Grote, T., Volpi, G., & Invernizzi, R. (2021). *'I See Something You Don't See'. A Computational Analysis of the Digital Services Act and the Digital Markets Act* [SSRN Scholarly Paper](ID 3780938). https://papers.ssrn.com/abstract=3780938

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3780938

https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3929276_code2064089.pdf?abstractid=3780938&mirid=1

Dictionary, C. (2022a). man. In *Cambridge Dictionary*. Retrieved 2022/04/14/17:05:18, from https://dictionary.cambridge.org/dictionary/english/man

Dictionary, C. (2022b). wing. In Retrieved 2022/04/14/14:49:36, from https://dictionary.cambridge.org/dictionary/english/wing

Elbow Method for optimal value of k in KMeans. (2019). *GeeksforGeeks*. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

Explainer: What's behind Thailand's protests? (2020, 2020/10/15/T05:23:14Z). *Reuters*. https://www.reuters.com/article/us-thailand-protests-reasons-explainer-idUSKBN2700IX

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), 201720347. https://doi.org/10.1073/pnas.1720347115

Hämäläinen, M., Patpong, P., Alnajjar, K., Partanen, N., & Rueter, J. (2021, 2021/11//). Detecting Depression in Thai Blog Posts: a Dataset and a Baseline.

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

How suitable is Twitter for monitoring policy developments? (2021). *Policy-Insider.AI*. https://policy-insider.ai/2021/08/23/how-suitable-is-twitter-for-monitoring-policy-developments/

Hu, R., Li, S., & Liang, S. (2019, 2019/07//). Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. ACL 2019,

HuggingFace. (2022, 2022/04/16/13:24:45). *Summary of the tokenizers*. https://huggingface.co/docs/transformers/tokenizer_summary

iLaw. (2021, 2021/01/06/). "ของอร่อยที่สุดก็คือ ภาษีประชาชน". https://www.facebook.com/photo/?fbid=10164861896145551&set=pcb.10164861922190551

Jin, X., & Han, J. (2010). K-Means Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 563-564). Springer US. https://doi.org/10.1007/978-0-387-30164-8_425

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/

Kam-Phaka. (2019, 2019/12/10/T10:22:32+00:00). คำ ผกา | ชังชาติทำไม. มติชนสุดสัปดาห์. https://www.matichonweekly.com/column/article_254411

Kaveh-Yazdy, F., & Zarifzadeh, S. (2021). Measuring Economic Policy Uncertainty Using an Unsupervised Word Embedding-based Method. *ArXiv*. https://doi.org/10.2139/SSRN.3845847

Kudo, T., & Richardson, J. (2018, 2018/11//). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. http://arxiv.org/abs/1907.11692

https://arxiv.org/abs/1907.11692

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). WangchanBERTa: Pretraining transformer-based Thai Language Models. https://arxiv.org/abs/2101.09635v2

https://arxiv.org/abs/2101.09635

McCarthy, A. D., Scharf, J., & Dore, G. M. D. (2021, 2021/11//). A Mixed-Methods Analysis of Western and Hong Kong–based Reporting on the 2019–2020 Protests. CLFL-EMNLP-LaTeCH-LaTeCHCLfL 2021,

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. http://arxiv.org/abs/1301.3781

https://arxiv.org/abs/1301.3781

Mob Data Thailand. (2022, 2022/01/11/). https://www.mobdatathailand.org

https://www.mobdatathailand.org/

Montariol, S., Allauzen, A., & Kitamoto, A. (2020, 2020/05//). Variations in Word Usage for the Financial Domain. COLING-FinNLP 2020,

Nelson, L. K. (2021). Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century U.S. South. *Poetics*, *88*, 101539. https://doi.org/10.1016/j.poetic.2021.101539 (Measure Mohr Culture)

Pennington, J., Socher, R., & Manning, C. (2014, 2014/10//). GloVe: Global Vectors for Word Representation. EMNLP 2014,

Schuster, M., & Nakajima, K. (2012, 2012). Japanese and korean voice search.

Stoltz, D. S., & Taylor, M. A. (2021). Cultural Cartography with Word Embeddings. *Poetics*, *88*, 101567. https://doi.org/10.1016/j.poetic.2021.101567

Tahmasbi, F., Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., & Zannettou, S. (2021, 2021/04/19/). "Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. Proceedings of the Web Conference 2021, New York, NY, USA.

thaidatapointscom. (2020). Twitter Analysis of the Thai Free Youth Protests. *Thaidatapoints*. https://www.thaidatapoints.com//post/twitter-analysis-of-the-thai-free-youth-protests

https://www.thaidatapoints.com/post/twitter-analysis-of-the-thai-free-youth-protests

Thairath. (2021, 2021/08/26/T05:08:00+07:00). ภาษาไทย (ใหม่) วันละคำ ความหมายของคำว่า "สลิ่ม". *www.thairath.co.th*. https://www.thairath.co.th/news/politic/2176568

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*. http://arxiv.org/abs/1706.03762

https://arxiv.org/abs/1706.03762

Vish. (2020). How Much Data Is Created Every Day in 2022? [You'll be shocked!]. *TechJury*. https://techjury.net/blog/how-much-data-is-created-every-day/

https://techjury.net/blog/how-much-data-is-created-every-day/#gref

VISTEC. (2021). WangchanBERTa โมเดลประมวลผลภาษาไทยที่ใหญ่และก้าวหน้าที่สุดในขณะนี้. *AIResearch.in.th*. https://medium.com/airesearch-in-th/wangchanberta-%E0%B9%82%E0%B8%A1%E0%B9%80%E0%B8%94%E0%B8%A5%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B8%A1%E0%B8%A7%E0%B8%A5%E0%B8%9C%E0%B8%A5%E0%B8%A0%E0%B8%B2%E0%B8%A9%E0%B8%B2%E0%B9%84%E0%B8%97%E0%B8%A2%E0%B8%97%E0%B8%B5%E0%B9%88%E0%B9%83%E0%B8%AB%E0%B8%8D%E0%B9%88%E0%B9%81%E0%B8%A5%E0%B8%B0%E0%B8%81%E0%B9%89%E0%B8%B2%E0%B8%A7%E0%B8%AB%E0%B8%99%E0%B9%89%E0%B8%B2%E0%B8%97%E0%B8%B5%E0%B9%88%E0%B8%AA%E0%B8%

B8%E0%B8%94%E0%B9%83%E0%B8%99%E0%B8%82%E0%B8%93%E0%B8%B0%E0%B8%99%E0%B8%B5%E0%B9%89-d920c27cd433

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., . . . Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*. http://arxiv.org/abs/1609.08144

https://arxiv.org/abs/1609.08144

APPENDIX

Similar word sets of all focus words

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
สิ่ง	หม้อ	ไตปลา	ไตปลา	ฉีดน้ำ	พยายามฆ่า	ฉีดน้ำ	จัดฉาก
แม่	ยูธ	ฮังเล	ฮังเล	ฉีดวัคซีน	บูด	ทุบ	ร้องเรียน
โลก	แผน	กะหรี่	ส้มกุ้ง	ขับรถชน	จัดฉาก	ฉีดวัคซีน	สกัด
น่า	แกนนำ	ส้มกุ้ง	เขียวหวาน	สกัด	แก๊ส	แก๊ส	ซ่อม
ค่า	แก	เขียวหวาน	น้ำแกง	ล้อม	โจมตี	ล้อม	แก๊ส
สามารถ	โอเค	เครื่องแกง	พริกขิง	ขยะแขยง	ทำงานให้	เดือด	โจมตี
น.	รับผิดชอบ	น้ำแกง	น้ำเงี้ยว	แก๊ส	สกัด	ซ่อม	ทุบ
ตั้งแต่	เตือน	แกงเขียวหวาน	บุ๊ง	บูด	ต้อน	เครื่องแบบ	ฉีดน้ำ
รอบ	สถานที่	น้ำพริก	ยูธ	อห	ทิพย์	สกัด	พระราชทาน
...	จัด	น้ำเงี้ยว	น้ำพริก	ไหม้	ไปแจ้งความ	จงใจ	บูด

Table 16: Similar word set of the focus word แกง “deceive” from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ผม	เจียว	ไตปลา	ไตปลา	ข้าว	ต้ม	ข้าว	ต้ม
ตรง	หอม	ฮังเล	ฮังเล	ผัก	แกงส้ม	อาหาร	ปลาร้า
out	ทอด	ส้มกุ้ง	ส้มกุ้ง	อาหาร	ข้าวต้ม	ผัก	แกงส้ม
ชวน	หม้อ	เขียวหวาน	เขียวหวาน	หมี่	น้ำพริก	ต้ม	น้ำพริก
เข้า	😋	กะหรี่	อ่อม	ก๋วยเตี๋ยว	ก๋วยเตี๋ยว	น้ำ	สุกี้
ชาติ	หั่น	อ่อม	น้ำแกง	ต้ม	ปลาทู	ก๋วยเตี๋ยว	ยำ
หน่อย	ปลา	น้ำแกง	น้ำเงี้ยว	น้ำ	ไข่	หมี่	ไข่เจียว
ใส่	ยยยย	แกงเขียวหวาน	พริกขิง	ขิง	ยำ	หมู	ไข่
เหตุ	นม	เครื่องแกง	พะแนง	หมู	กะทิ	บัตร	ผัด
ข้าง	ไก่	น้ำพริก	เครื่องแกง	เย็น	ผัด	ขิง	กะทิ

Table 17: Similar word set of the focus word แกง “deceive” from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ยุค	ชัง	ชนชาติ	ปชช	ชิง	ชิง	ชิง	ชิง
สมอง	ชั่ว	ชน	หัวควย	สัตว์	ชาต	สัตว์	สัตว์
น.	ทหารเหี้ย	ความดื้อ	ใช้เผด็จการ	ชาต	สัตว์	นรก	ชั่ว
ถูก	แผ่นดิน	เพณี	สถุน	นรก	รักชาติ	ชาต	ประเทศ
เด็ก	ม็อบสามกีบ	ซี้ซั้ว	ทั้งๆที่	เลว	เลว	ประเทศ	ชาต
มวล	ปรสิต	นายสั่ง	ทำร้ายปชช	กรรม	ประเทศ	รักชาติ	เลว
4	แท้	(ชทพ)	สัส	ระยำ	ชั่ว	เลว	กรรม
เล่น	อุบาทว์	“ยิ	ประนาม	ประเทศ	กรรม	กรรม	รักชาติ
โดย	ตระกูล	คนป่าเถื่อน	ฆาตรกร	พวก	นรก	ชั่ว	บัด
ห้าม	ปล้น	ลม ๆ แล้ง ๆ	ซี้ซั้ว	รักชาติ	เพื่อชาติ	พวก	นรก

Table 18: Similar word set of the focus word ชาติ “nation” from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
การ	ชัง	ชนชาติ	ชนชาติ	ของชาติ	ของชาติ	ประเทศ	ของชาติ
เข้า	แผ่นดิน	(ชทพ)	ประเทศไทย	ประเทศ	ประเทศ	ของชาติ	ประเทศ
ฟ้า	ไทย	เพณี	จัญไร	โลก	รักชาติ	ฝ่าย	รักชาติ
จำนวน	ต่าง	คนป่าเถื่อน	ชาติศาสน์	รักชาติ	ชาติไทย	รัฐ	แผ่นดิน
ไว	ธง	ความเกื้อกูล	มักใหญ่ใฝ่สูง	รัฐ	แผ่นดิน	โลก	ชาติไทย
โดย	วัฒนธรรม	มักใหญ่ใฝ่สูง	แหล่ะ	ชาติไทย	รัฐ	ค่าย	รัฐ
ทำลาย	พระมหากษัตริย์	ความดื้อ	กฎหมาย	ฝ่าย	ค่าย	รักชาติ	ค่าย
อะ	กษัตริย์	กฎหมาย	ล้มเจ้า	ไทย	สถาบัน	โรค	โรค
ปาก	ศาสนา	(ทษช)	เห่อเหิม	ค่าย	ไทย	รัฐบาล	เมือง
ง่าย	ประเทศ	ซี้ซั้ว	ปชช	กรรม	ต่างชาติ	กรรม	เพศ

Table 19: Similar word set of the focus word ชาติ “nation” from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ฝึก	ทำร้าย	แม่ง	เป็นฆาตกร	ของประชาชน	ของประชาชน	ของประชาชน	ชาวบ้าน
ไว้	ประชาชนอะ	“ลง	ทำร้ายปชช	ให้ประชาชน	ให้ประชาชน	คนไทย	ทุกคน
วาง	ปชช.	ๆ	ประชาชนอะ	กับประชาชน	ชาวบ้าน	ทุกคน	ของประชาชน
ครั้ง	ตอบแทน	ระงับใจ	อิพวก	ปชช	กับประชาชน	ให้ประชาชน	พวกเรา
ราษฎร	ประชาชนอ่ะ	และ	ประนาม	ทุกคน	จากประชาชน	ประเทศ	คนอื่น
หมา	ปปช.	คนสารเลว	ใช้เผด็จการ	คนไทย	ประชา	กับประชาชน	ให้ประชาชน
ล้ม	ศัตรู	สมเพช	ตีนเผด็จการ	จากประชาชน	ปชช	ตัวเอง	คนไทย
ต้อง	ใช่	หัวควย	ข้าเผด็จการ	ประชา	พวกเรา	ชาวบ้าน	ประเทศ
เผด็จการ	ปกป้อง	แตกสามัคคี	ประชาชนอ่ะ	ชาวบ้าน	ทุกคน	กฎหมาย	ประชา
เลี้ยง	หรอ?	เอาตาย	ปะวะ	พวกเรา	คนไทย	คนอื่น	ตัวเอง

Table 20: Similar word set of the focus word ประชาชน “people” from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ห้าม	จัดการ	อัลลอฮ์)	ปชช	ทุกคน	ของประชาชน	ชาวบ้าน	ของประชาชน
5	เด็ดขาด	แตกสามัคคี	ปชช.	ชาวบ้าน	กับประชาชน	ทุกคน	ชาวบ้าน
55555555	ปชช	มักใหญ่ใฝ่สูง	บริหารจัดการ	คนไทย	ให้ประชาชน	ประเทศ	กับประชาชน
การ	กฎ	ปชช	เหี้ยไร	ประเทศ	ชาวบ้าน	คนไทย	ให้ประชาชน
อะ	เพื่อ	สิ่งเร้นลับ	สามกีบ	เด็ก	ทุกคน	สังคม	ลูกค้า
รวม	สมควร	ความทุจริต	ไอ้เหี้ย	รัฐบาล	จากประชาชน	รัฐบาล	ทุกคน
สร้าง	เป้าหมาย	“อัลลอฮ์	ชั่วๆ	สังคม	ลูกค้า	เด็ก	รัฐบาล
%	กระบวนการ	คิดคด	กฎหมาย	อยู่	คนไทย	คนอื่น	ประเทศ
เช่น	ล้มเหลว	จับปลาสองมือ	เฟคนิวส์	กับประชาชน	ประชา	ตัวเอง	คนไทย
-	อัยการศึก	พสก	แม่ม	ผู้	ประเทศ	อยู่	ตำรวจ

Table 21: Similar word set of the focus word ประชาชน “people” from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
น.	ตกต่ำ	สมองทึบ	ประเทศเหี้ย	ประเทศไทย	ประเทศไทย	โลก	ประชาชน
ตร.	ประเทศชาติ	อวดเก่ง	รัฐบาลเหี้ย	ของประเทศ	ของประเทศ	ประชาชน	เกิด
ไอ	เจริญ	เอาตาย	ประเทศไทย	โลก	โลก	บ้าน	ตัวเอง
รอบ	ประเทศเหี้ย	คนป่าเถื่อน	บริหารจัดการ	ปี	บ้านเมือง	ปี	ชีวิต
ใน	เน่าเฟะ	ทำร้ายปชช	ห่าไร	บ้านเมือง	ประชาชน	อยู่	โลก
ตู้	ปท	สถุล	มึง	ประชาชน	ชีวิต	ชีวิต	ไทย
ปัญหา	ถ่วง	โดนดี	สาปส่ง	ไทย	ปี	ไทย	บ้านเมือง
กษัตริย์	ท้อแท้	สมบัติผู้ดี	ประเทศเฮงซวย	ในประเทศ	ประเทศนี้	ตัวเอง	เมือง
กอง	ซวย	เล่นสกปรก	อิกนอร์	หมด	ไทย	ประเทศไทย	ประเทศไทย
เจอ	พัฒนา	ผิดเพศ	เป็นทุกข์เป็นร้อน	ชีวิต	บ้าน	หมด	ระบบ

Table 22: Similar word set of the focus word ประเทศ “country” from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
สิ่ง	พัฒนา	เมกา	ประเทศไทย	ปี	ของประเทศ	ประชาชน	ปี
อะ	ต่าง	ปัญา	เมืองไทย	รัฐบาล	ปี	ปี	รัฐบาล
ทำลาย	เมืองไทย	clmvt	ปชช	ประชาชน	ไทย	รัฐบาล	ประชาชน
กาย	ไทย	แอสตรา	บริหารจัดการ	ไทย	ชาติ	บริษัท	พรรค
มอง	ปท	อีรุงตุงนัง	โซเชี่ยล	ชาติ	โลก	ร้าน	ชาติ
โดย	ร่ำรวย	๔.๐	บริจา	โลก	ประชาชน	สังคม	ไทย
ยืน	มหาอำนาจ	🌻	วัคซีนฯ	บริษัท	ในประเทศ	บ้าน	โลก
การ	ประเทศไทย	ในประเทศ	บุลลี่	ดี	รัฐบาล	โลก	เมือง
กี่	โลก	แมว้า	เมกา	ของประเทศ	ประเทศไทย	ชีวิต	ทีม
ม็อบ	เพื่อนบ้าน	"ρ	ห่าไร	ทาง	จังหวัด	ไทย	ร้าน

Table 23: Similar word set of the focus word ประเทศ “country” from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ตั้งใจ	จ่าย	เสียภาษี	ทำร้ายปชช	เงิน	เงินภาษี	เสียภาษี	ประชาชน
ปล่อย	เดือน	เก็บภาษี	ประชาชนอ่ะ	เสียภาษี	เสียภาษี	เงิน	เกิด
สิงหา	เงิน	ภาษีการค้า	เงิน	เงินเดือน	ใช้เงิน	เงินเดือน	ตัวเอง
กษัตริย์	ตอบแทน	ภาษีอากร	ปชช	จ่ายเงิน	เงินเดือน	จ่ายเงิน	ชีวิต
ข้อ	แดก	ภาษีรถยนต์	แหล่ะ	ประชาชน	เงิน	จ่าย	โลก
ช่วง	แดก	ภาษีรายได้	กินดี	สวัสดิการ	งบ	ประชาชน	ไทย
ถือ	ซื้อ	ภาษีเงินได้	ประชาชนอะ	เลือด	เอาเงิน	จ่ายภาษี	บ้านเมือง
สามารถ	สวัสดิการ	ภาษีโรงเรือน	ปราบม็อบ	จ่าย	จ่ายภาษี	สวัสดิการ	เมือง
4	เนรคุณ	ภาษีสรรพสามิต	หัวควย	งบ	ประชาชน	ข้าว	ประเทศไทย
ปั่น	ตัง	earmarked	ไอเหี้ย	ข้าว	งบประมาณ	มาจาก	ระบบ

Table 24: Similar word set of the focus word ภาษี "tax" from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
โดย	จ่าย	เสียภาษี	เสียภาษี	งบ	เสียภาษี	เงิน	เสียภาษี
ข่าวสาร	เงิน	เก็บภาษี	ภาษีการค้า	จ่าย	เก็บภาษี	จ่าย	เก็บภาษี
งบ	หนี้	ภาษีอากร	ภาษีอากร	เงิน	vat	งบ	เงิน
เจ้า	จ้าง	ภาษีการค้า	ภาษีรายได้	vat	รายได้	ข้าว	งบ
หลับ	เยียวยา	ภาษีรถยนต์	เงิน	เก็บภาษี	งบ	บัตร	ประกัน
อะ	แพง	ภาษีรายได้	ภาษีเงินได้	ค่า	เงิน	ทหาร	รายได้
จับ	สวัสดิการ	ภาษีเงินได้	เก็บภาษี	ปรับ	จ่าย	ปรับ	จ่าย
โคตร	คุ้ม	ภาษีสรรพสามิต	earmarked	รัฐ	ประกัน	เก็บภาษี	vat
การ	ประกัน	ภาษีโรงเรือน	เจ้าภาษี	ประกัน	ลดหย่อน	ประชาชน	เลือด
ปี	กำไร	ภาษีมูลค่าเพิ่ม	ภาษีมูลค่าเพิ่ม	เสีย	สรรพากร	เสีย	โกง

Table 25: Similar word set of the focus word ภาษี "tax" from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
วัน	ธรรมเนียบ รัฐบาล	ม็อป	ม๊อบ	ม๊อบ	ม๊อบ	ม๊อบ	ชุมนุม
เปิด	วัน	vileda	ม็อป	ผู้ชุมนุม	ชุมนุม	ชุมนุม	ม๊อบ
ฝาก	,rt	ม๊อบ	ม็อบ26	ชุมนุม	การ์ด	การชุมนุม	ประท้วง
ล้าน	Please	newsponge	vileda	การชุมนุม	คาร์	ผู้ชุมนุม	แกนนำ
)		pva	newsponge	การ์ด	แกนนำ	แกนนำ	การชุมนุม
เจอ	สาธารณะสุข	swash	Retweet	ล้อม	ผู้ชุมนุม	แยก	สถานการณ์
เพิ่ม	Retweet	supercat	ม้อบ	แกนนำ	หาดใหญ่	การ์ด	ขบวน
หน่วย	หาดใหญ่	ไม้ถูพื้น	supercat	รัฐประหาร	มวลชน	ประท้วง	ไฟไหม้
แกนนำ	ธรรมเนียบ	be-man	ธรรมเนียบ	แยก	อาชีวะ	โจร	การ์ด
เรา	💢	ธรรมเนียบ	pva	แก๊สน้ำตา	การจับกุม	ล้อม	แยก

Table 26: Similar word set of the focus word ม็อบ "protest" from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
จาก	ม๊อบ	ม็อป	ม็อป	ม๊อบ	ม๊อบ	ม๊อบ	ชุมนุม
ขอ	ชุมนุม	vileda	ม๊อบ	ชุมนุม	ชุมนุม	ชุมนุม	ม๊อบ
ผล	ป่วน	ม๊อบ	vileda	แกนนำ	แกนนำ	แกนนำ	แกนนำ
สิ่ง	สลาย	newsponge	newsponge	การชุมนุม	การชุมนุม	การชุมนุม	การชุมนุม
ขนาด	ตำรวจ	pva	pva	ประท้วง	ประท้วง	ประท้วง	ประท้วง
รถ	เผา	ไม้ถูพื้น	supercat	ขบวนการ	ผู้ชุมนุม	สลิ่ม	เสื้อแดง
โดย	ฝูง	supercat	สามกีบ	ป่วน	เสื้อแดง	ตำรวจ	แก๊ง
2	ดินแดง	swash	ไอ้เหี้ย	สลิ่ม	แก๊ง	ผู้ชุมนุม	ผู้ชุมนุม
อะ	คฝ	be-man	swash	มวลชน	เด็กแว้น	มวลชน	แท็ก
ยุค	แกนนำ	ncl	ประนาม	ตำรวจ	ฝูงชน	ป่วน	ฝูงชน

Table 27: Similar word set of the focus word ม็อบ "protest" from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
สมอง	เฮงซวย	ประเทศไทย	แหล่ะ	รัฐ	รัฐบาลนี้	ตำรวจ	รัฐบาลนี้
กษัตริย์	ส้น	ม็อบ10	รัฐบาลเหี้ย	ของรัฐบาล	ของรัฐบาล	ของรัฐบาล	รัฐ
36	ตีน	ทำร้ายปชช	ประยุทธ	ตำรวจ	รัฐบาลไทย	รัฐ	ของรัฐบาล
จุด	ล้มเหลว	สัส	บริหารจัดการ	รัฐบาลไทย	กับรัฐบาล	กับรัฐบาล	รัฐบาลไทย
โลก	ม็อบส้น	"ธนกร"	ม็อบ2กรกฎา	ฝ่ายค้าน	และรัฐบาล	รัฐบาลนี้	ตำรวจ
หน่วย	ประเทศส้น	ปชช	รัฐบาลฆาตกร	จากรัฐบาล	รัฐ	นายกฯ	นายก
เกลือ	รบ.	ๆ	ม็อบ18	นายก	จากรัฐบาล	อห	กับรัฐบาล
กันยา	ส้นตีน	เล่นไม่ซื่อ	ชห	อห	ที่รัฐบาล	ศาล	และรัฐบาล
ยุค	เชียร์	เล่นสกปรก	มึง	กับรัฐบาล	ให้รัฐบาล	จากรัฐบาล	ที่รัฐบาล
ด่าน	อีรัฐบาล	กฎหมาย	ประเทศส้น	วัคซีน	นายกฯ	ประเทศชาติ	นายกฯ

Table 28: Similar word set of the focus word รัฐบาล "government" from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
อ่าน	รบ.	กดทับ	ปชช	รัฐ	รัฐ	รัฐ	รัฐ
ลุง	รบ	ซึ่ง ๆ หน้า	โน่นนี่	ของรัฐบาล	ของรัฐบาล	ของรัฐบาล	ของรัฐบาล
ภูมิ	ห่วย	" "	ล้มเจ้า	นายก	รัฐบาลนี้	ประเทศ	นายก
ทุก	จัดการ	"ธนกร"	ทำไร	ฝ่ายค้าน	ภาครัฐ	นายก	ภาครัฐ
ข้าว	ประชาชน	ประชาธิปไตย	บริหารจัดการ	ประเทศ	ฝ่ายค้าน	ประชาชน	ฝ่ายค้าน
ชาว	ฝ่าย	ตัดช่องทาง	สามกีบ	ประชาชน	เป็นรัฐบาล	ตำรวจ	ตำรวจ
ได้	สส	เล่นไม่ซื่อ	บลา	รบ	รัฐบาลไทย	โครงการ	รัฐบาลนี้
เสร็จ	เฮงซวย	เฟคนิวส์	ขั๋ว	โครงการ	นายก	ทหาร	ประเทศ
จาก	โจมตี	ปึ้ง'	"ธนกร"	ตำรวจ	ฝ่ายรัฐบาล	บริษัท	ประชาชน
ไว้	เด็ดขาด	ปชช	แม่ม	ของรัฐ	กับรัฐบาล	ตัวเอง	โครงการ

Table 29: Similar word set of the focus word รัฐบาล "government" from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
รอบ	สิทธิ์	ประณีประนอม	ผิดหรอ	ประท้วง	ประท้วง	ประท้วง	ประท้วง
น.	พื้นฐาน	ผินหลังให้	ทั้งๆที่	ต้องการ	ทวง	ขอความ	ต่อสู้
เอง	แค่	ผ่อนสั้นผ่อนยาว	มึง	ต่อสู้	ต่อต้าน	ต้องการ	ต่อต้าน
สิ่ง	สิ่ง	สำรวมใจ	ทำร้ายปชช	ขอความ	ต่อสู้	ต่อสู้	ทวง
เล่น	เป็นธรรม	ศักเคียส	กิบลัต	ร้องขอ	ร้องขอ	ร้องขอ	ทวงคืน
แตก	ประชาธิปไตย	ความร้อนรน	อิกนอร์	ปกป้อง	ทวงคืน	สู้เพื่อ	อยากได้
เด็ก	ถูกต้อง	medicare	กินดี	สู้เพื่อ	ขับไล่	ปกป้อง	แสดงออก
หน่วย	ด้วยซ้ำ	ayah	แหล่ะ	ชุมนุม	แสดงจุดยืน	ชุมนุม	ปกป้อง
ด่าน	รับฟัง	ทุ่มเถียง	โอนอ่อนผ่อนตาม	อยากได้	แสดงออก	เรียก	ต้องการ
กลุ่ม	ผิดหรอ	เข้าร่องเข้ารอย	” “	เรียก	สู้เพื่อ	แสดงความ	สู้เพื่อ

Table 30: Similar word set of the focus word เรียกร้อง “demand/call for” from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ลด	สิทธิ	ประณีประนอม	บุลลี่	แอนตี้	ประท้วง	ประท้วง	ประท้วง
ภูมิ	อ้าง	ผินหลังให้	ปชช	เสรีภาพ	ต่อต้าน	เสรีภาพ	ทวง
ปล่อย	ละเมิด	ผ่อนสั้นผ่อนยาว	บูลลี่	นม	ต้องการ	โหน	ต่อต้าน
เห็น	กระบวนการ	ความร้อนรน	สามกีบ	เห็นต่าง	ทวง	ทวง	กดดัน
เข้า	โจมตี	สำรวมใจ	หลับหูหลับ	ขยัน	ถามหา	เคลม	ถามหา
เสียใจ	เสรีภาพ	พสก	กิบลัต	เสียงดัง	ร้องขอ	ปะทะ	รณรงค์
ใหม่	ต่อต้าน	ใฝ่ต่ำ	มาดราม่า	หลุด	กดดัน	อยากได้	ต้องการ
เดิม	มนุษยชน	ทุ่มเถียง	กฎหมาย	ร้านค้า	รณรงค์	ประกาศ	สนับสนุน
เฉพาะ	ยัดเยียด	ศักเคียส	บังคับใจ	io	โหยหา	งอแง	ขับไล่
รถ	กฎหมาย	คาดหมาย	” “	ประท้วง	โวยวาย	ต้องการ	งอแง

Table 31: Similar word set of the focus word เรียกร้อง “demand/call for” from opposition corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
ข้อ	เบิกเนตร	ดัดจริต	ปิดหู	บูด	ไอโอ	แดง	เสื้อแดง
น.	กะลา	ไล่ส่ง	อีตู่	ตู่	เสื้อแดง	เสื้อแดง	ไอโอ
ถ่าย	แหกตา	สมองทึบ	รัฐบาลเหี้ย	แดง	io	ตู่	io
ฝึก	บูด	สามหาว	ประเทศส้น	เสื้อแดง	นกหวีด	ม็อบ	ตู่
ฝาก	โทรทัศน์	ขนมเปียกปูน	จัญไร	เพนกวิน	บูด	ผู้ชุมนุม	ติ่ง
เจอ	มัว	เฉิ่ม	สมองทึบ	อีเหี้ย	ตู่	แกนนำ	ควาย
ถูก	ลืมหูลืม	ปลาท่องโก๋	ขี้ขโมย	ขี้ข้า	ติ่ง	ควาย	ไดโนเสาร์
พร้อม	หลับหูหลับ	ซาหริ่ม	ปิดตา	ผู้ชุมนุม	กะทิ	ขี้ข้า	อีเหี้ย
ตร.	ก้	เง้า	อีสลิ่ม	รัฐบาล	กปปส	คนโง่	แกนนำ
รู้	รำคาญ	พองขน	ปั่นแท็ก	ม๊อบ	ควาย	ประยุทธ์	แดง

Table 32: Similar word set of the focus word สลิ่ม "Sa-lim" from protester corpus

Un-pre-trained static word embedding		Pre-trained static word embedding		Contextualized word embedding		Finetuned contextualized word embedding	
Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset	Small size dataset	Large size dataset
อะ	ทัวร์	ดัดจริต	บูลลี่	ติ่ง	ติ่ง	ติ่ง	ติ่ง
สนุก	อวย	ไล่ส่ง	สามกีบ	เสื้อแดง	เสื้อแดง	เสื้อแดง	เสื้อแดง
มือ	ด้อม	สามหาว	สำออย	ควาย	สาวก	ควาย	ควาย
สี	กรู	ซาหริ่ม	5555555555555	แดง	ควาย	ม็อบ	ส้ม
เร็ว	อาย	สมองทึบ	สมองทึบ	ส้ม	ส้ม	แดง	แดง
หรือ	แท็ก	เฉิ่ม	" "ผม	ม็อบ	แดง	ส้ม	สาวก
วัย	ผี	แหย	เป็นมิน	สาวก	io	เอกชน	ม็อบ
,	โควท	ขนมเปียกปูน	ประนาม	ตู่	กะเทย	ทหาร	กะเทย
AZ	งง	ขี้ขโมย	5555555555	อคติ	ม็อบ	แกนนำ	แกนนำ
สิ่ง	เถียง	...?	หนีหน้า	แกนนำ	ไอโอ	สาวก	ตู่

Table 33: Similar word set of the focus word สลิ่ม "Sa-lim" from oppostion corpus

Proportion of corpus in the clusters of contextualized word embedding

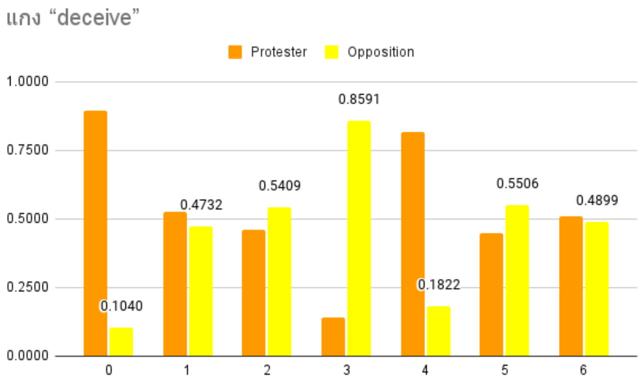


Figure 29: Proportion of corpus in the clusters of แกง "deceive"

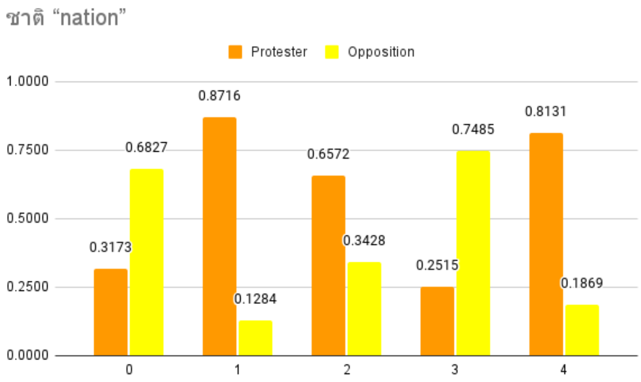


Figure 30: Proportion of corpus in the clusters of ชาติ "nation"

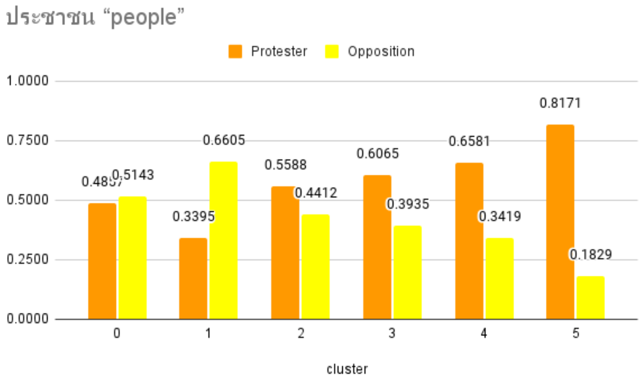


Figure 31: Proportion of corpus in the clusters of ประชาชน "people"

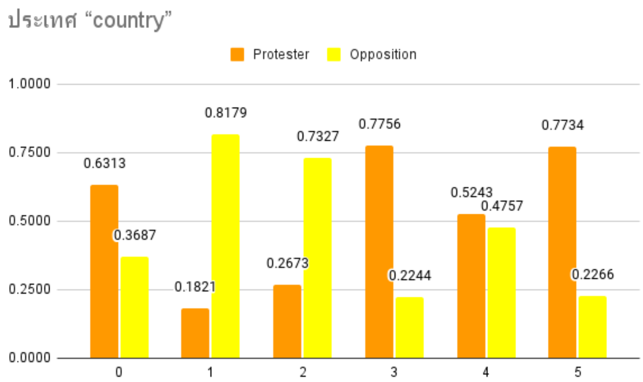


Figure 32: Proportion of corpus in the clusters of ประเทศ "country"

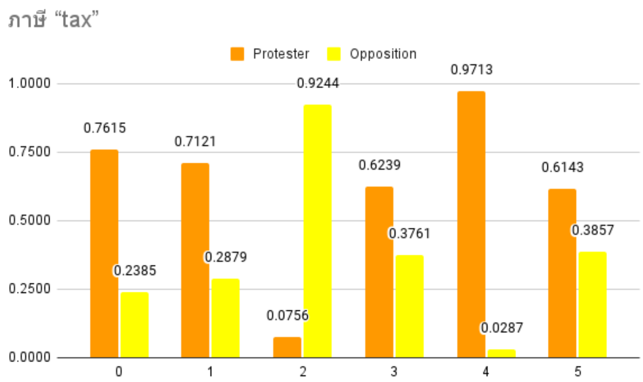


Figure 33: Proportion of corpus in the clusters of ภาษี "tax"

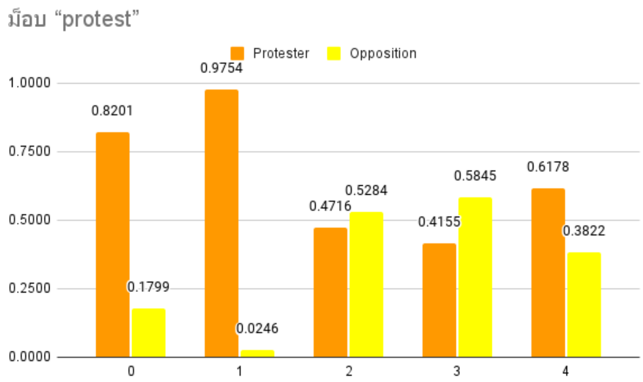


Figure 34: Proportion of corpus in the clusters of ม็อบ "protest"

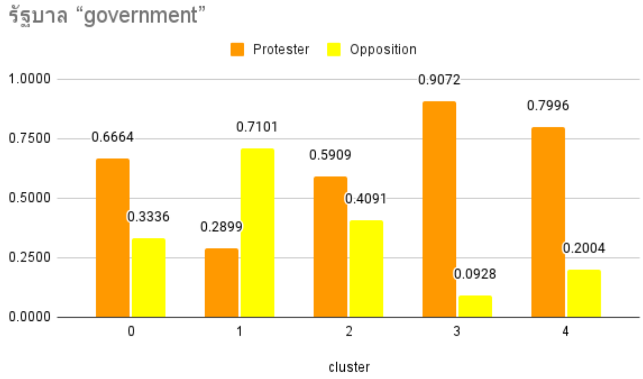


Figure 35: Proportion of corpus in the clusters of รัฐบาล “government”

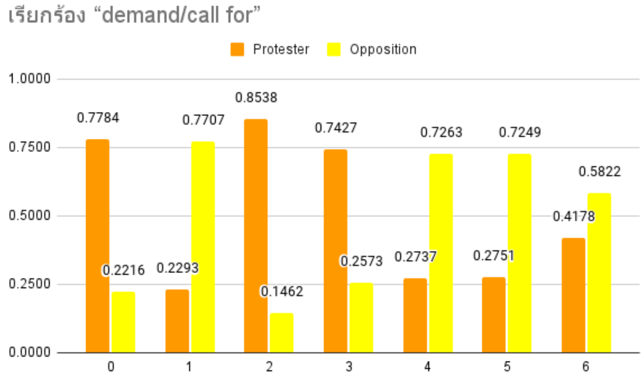


Figure 36: Proportion of corpus in the clusters of เรียกร้อง “demand/call for”

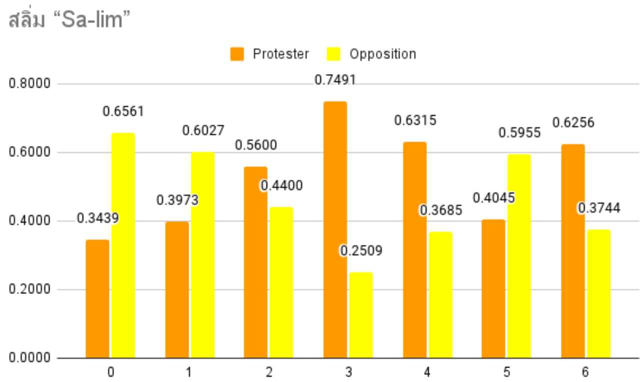


Figure 37: Proportion of corpus in the clusters of สลิ่ม “Sa-lim”

VITA

NAME	Wassakorn Sarakul
DATE OF BIRTH	1 May 1995
PLACE OF BIRTH	Chanthaburi, Thailand
INSTITUTIONS ATTENDED	Chulalongkorn University