SENTIMENT ANALYSIS FOR PHARMACEUTICAL COMPANY FROM SOCIAL MEDIA USING ADAPTIVE COMPRESSION (ADACOMP) WITH RANDOM UNDER SAMPLE (RUS) AND SYNTHETIC MINORITY OVER-SAMPLING (SMOTE)

Pamungkas Setyo Wibowo[1*]; Andry Chowanda[2]

Computer Science Department
Binus Graduate Program – Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia, 11380
pamungkas.wibowo@binus.ac.id [1*], achowanda@binus.edu[2]

Abstract—Pharmaceutical company has become the most highlight company across the world lately because of the pandemic. Despite of the high demand market in pharmaceutical company, about 94% of large company across the world having difficulty in their supply chain that indirectly affect their services. The purpose of this research is to compare word embedding with compression model by doing sentiment analysis about the entity to find the best model that give better accuracy rates based on the opinion of Twitter, Instagram and Youtube, as they are the largest platform that its many users to express their opinions about an individual or an instance. Data is retrieved from Twitter, Instagram and Youtube using the R-Studio application by utilizing their API library, then preprocessing and stored in a database. Next step is labeling the data and then train the data using word2Vec and LSTM, GloVe and LSTM and lastly using Adaptive Compression (adaComp) to compress the both model word embedding. Unfortunately, we got imbalanced dataset after labeling process, so we add sampling technique to sampling the dataset using Random Under Sample (RUS) and Synthetic Minority Over-sampling Technique (SMOTE). After the data are trained and tested, the results will be evaluated using Confusion Matrix to get the best Accuracy. With several models that have been carried out,applying adaComp is proven to increase accuracy. In the Word2Vec word embedding with LSTM model, applying adaComp increasing its accuracy from 77% to 81%.

Keywords: Pharmaceutical, Sentiment Analysis, AdaComp, Word2Vec, GloVe.

Intisari—Perusahaan farmasi telah menjadi perusahaan yang paling menonjol di seluruh dunia akhir-akhir ini karena pandemi. Meskipun permintaan pasar pada perusahaan farmasi sedang tinggi, sekitar 94% perusahaan besar di seluruh dunia mengalami kesulitan dalam memenuhi rantai pasokan mereka yang secara tidak langsung mempengaruhi layanan mereka. Tujuan dari penelitian ini adalah untuk mengetahui opini publik tentang Entitas Perusahaan Farmasi apakah positif, netral atau negatif dengan melakukan analisis sentimen tentang entitas itu sendiri berdasarkan pendapat Twitter, Instagram dan Youtube, dikarenakan platform tersebut adalah yang memiliki banyak penggunanya untuk mengekspresikan pendapat mereka tentang individu atau perusahaan. Data diambil dari Twitter, Instagram dan Youtube menggunakan aplikasi R-Studio dengan memanfaatkan library API-nya, kemudian dilakukan preprocessing dan disimpan dalam database. Langkah selanjutnya adalah pelabelan data dan kemudian melatih data menggunakan word2Vec dan LSTM, GloVe dan LSTM dan terakhir adalah mengaplikasikan adaptive compression(adaComp) pada word embedding. Sayangnya, kami mendapatkan dataset yang tidak seimbang setelah proses pelabelan, jadi kami menambahkan teknik sampling untuk mengambil sampel dataset menggunakan Random Under Sample (RUS) dan Synthetic Minority Over-sampling (SMOTE). Setelah data dilatih dan diuji, hasilnya akan dievaluasi menggunakan Confusion Matrix untuk mendapatkan Akurasi terbaik. Dengan beberapa model yang telah dilakukan, penerapan adaComp terbukti meningkatkan akurasi. Pada Word2Vec word embedding dengan model LSTM,

penerapan adaComp meningkatkan akurasinya dari 77% menjadi 81%.
Kata Kunci: Farmasi, Analisis Sentimen, AdaComp, Word2Vec, GloVe.

INTRODUCTION

Social media has become our daily needs and cannot be separated from human activity (Fabris et al. 2020). Even in situations of social uncertainty such as lockdown decisions and pandemics, the need for the use of social media is increasing (Størdal et al. 2021). Social media users is 59% of Indonesian Internet users, counting to 160 million Users (Kemp 2020). Some of the platforms that are widely used is Twitter, Instagram and Youtube where the number of users is 27% of the number of social media users in Indonesia(Muhammad, Kusumaningrum, and Wibowo 2021). Nowadays, social media is not only restricted to communicate with its own member but there is some official company that using social media to retrieve feedback from its user(Jiao, Veiga, and Walther 2020; Robiady, Windasari, and Nita 2020).

Sentiment analysis or opinion mining is a process of understanding, extracting and processing textual data automatically to get sentiment information contained in an opinion sentence (Liu, Shin, and Burns 2019). This is important for companies to be able to find out the value of opinions formed on social media (Rasool, Shah, and Islam 2020) so that it is hoped that the company can take the right steps in determining the best strategy. Lately, pharmaceutical companies are experiencing tremendous impact because of the effects of the pandemic as the need for medicine is the priority in every aspect(Prasad and Bodhe 2012). However, in addition to the extraordinary demand for drugs and supplements, many difficulties occurred, especially in the material supply chain for product manufacturing that impact about 94% to large company around the world (Chowdhury et al. 2021) . To fulfil this high demand on the market companies need to make extra effort that impact their services to society. For this reason there is high number of public opinion in social media had been created(De, Pandey, and Pal 2020).

The use of deep learning to conduct sentiment analysis can be a solution to these problems (Alam et al. 2020). In previous research sentiment analysis was only carried out on one social media. Meanwhile, according to statistical calculations carried out by statcounters, the status of use of social media in Indonesia is divided into many platforms. Generic word embedding such as GloVe and word2Vec which have been pre-trained have shown tremendous success when used, however there are many applications that use specialized vocabulary domains and the relatively small amount of data is not optimal (Sarma, Liang, and Sethares 2018).

Topic compression word embedding in sentiment analysis is to compress for each word in word embedding before used for sentiment analysis. For the last five years, many topics about sentiment analysis have been presented and the latest approaches about it is using compression in word embedding to support sentiment analysis. The use of adaptive compression is based on Gumbel-Softmax algorithm (Kim, Kim, and Lee 2020) by changing the concept of word embedding to adaptiv and learns to compress embedding words that take into account the downtime of the task. Yet for this last research is not applied to analysis sentiment, and most of the proposed topic segment in sentiment analysis are used for English language. The most known word embedding to support that are Word2Vec or GloVe which are considered as generic word embedding.

For example, works deal with indonesian sentiment analysis such as the work of Nawangsari et al (Nawangsari, Kusumaningrum, and Wibowo 2019). They conducted a study on word2vec by comparing the existing models and giving the best model result is Skip-Gram with an accuracy rate of 92.377%. On the other hand, Bagheri et al (Bagheri, Saraee, and Jong 2013)conducted a similar research on a dataset of customer reviews of electronic products to determine their value on customer satisfaction. This study give an accuracy value of 84.5% has been obtained and has been effective in seeing the level of customer satisfaction.

Based on these related works in the field of sentiment analysis using compression method, we notice that only single platform social media is used, especially the references of sentiment analysis that use compression method. However, the application of compression with the adaptive model is relatively new research. Previously there had been research related to adaptive compression of word embedding, but in this study the author will apply adaptive word embedding with the Indonesian language corpus with research subjects in pharmaceutical companies. Furthermore, we want to conduct research using various data sources not only from one social media and apply the adaptive compression to the Word2Vec and GloVe models.

Hence in this paper, we will study Word2Vec and GloVe with adaptive compression to determine which one is the most efficient method that give the best improvement in accuracy. Moreover, this study will be done with multi platform social media such as Twitter, Instagram and Youtube to improve the dataset due to pharmaceutical company entities are rarely found on social media. Furthermore we use LSTM processed layer to conduct sentiment

analysis and exploits the bilingual aspect of these methods by focusing on two languages : English and Indonesian.

This paper is organized as follows : Section 2 presents related works in the field of sentiment analysis; Section 3 explain about proposed method and describes Word2Vec, GloVe and Adptive compression (AdaComp); Section 4, experimental result and discussion are reported; The conclusion and future woek are presented in section 5.

MATERIALS AND METHODS

First, data will be collected on several social media which is Twitter, Youtube and Instagram. The data taken from social media contains the keyword of Entity that used as subject of this study. The process of retrieving data uses the API provided by social media developers and forum developer sites such as developers google and github. After the data is obtained, the data will be stored in the PostgreSQL database.

The data that has been stored in the database is raw data where there are still many invalid words, so pre processing will be carried out to clean the data and will be processed to label per row of data using lexicon, alternatively the data will be labelled manual with subjective appraisal from the annotator. After that the data will be entered in CSV and will be used as a data model for the Train and Test of the Learning algorithm in using several word embedding models compared to one adaptive model.

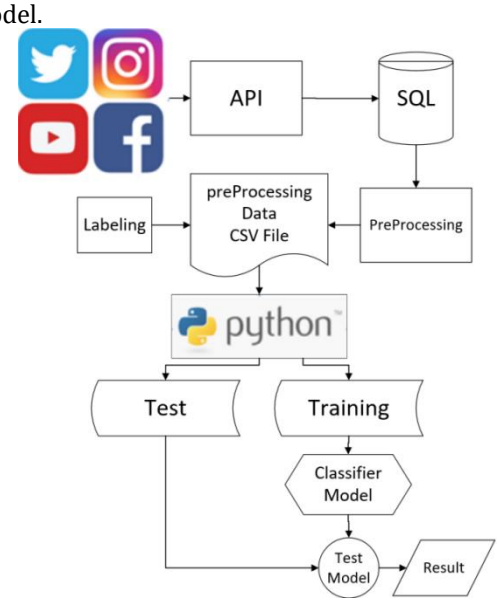

Figure.1 Methodology

a. Data Crawling

Data from Twitter will be collected through the Twitter API, Instagram API and Youtube API using R-Studio software. The data collected are about Entity referred by those social media users from August 1, 2017 to Febuary 15, 2021. After the data is obtained, the data is then stored in the PostgreSQL Database. But not all parts of the data are stored in a database. Information needed for research is as follows:

- ID: The ID for each Tweet, Instagram post and Youtube post that exists, originating from each social media.
- Content: Tweet, Instagram post and Youtube post from Twitter, Instagram and Youtube about Entity.
- Date: The date a post was created

From the data obtained a total of 33,864 data which is a combination of 3 data obtained from 3 annotators into a dataset. The dataset that will be used in this study was built by the author with details of 19,449 data from social media Instagram, 13213 data from social media Twitter and the last 1200 data from social media YouTube.

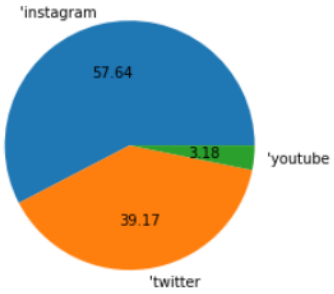

Figure.2 Percentage dataset source

b. PreProcessing

PreProcessing is a step to clean up data for sentiment analysis. The first PreProcessing steps is case folding, this step aims to turn all words into lowercase letters. The aim is to avoid case sensitive when matching words with a dictionary. An example is the change in the word 'Slow' to 'slow'. Second, normalization to remove the link in the post on social media, because the link is not part of the analysis. After that, we do data cleansing to removing characters other than letters, such as punctuation and symbols. Then we remove stopwords to eliminate words that are considered to have no meaning. Finally, tokenization is done by separating each word into one separate part. Separation of these words is done by cutting sentences based on spaces so that later can be made a vocabulary based on unique words contained in the text.

Vol. 19, No. 2 September 2022 | DOI https://doi.org/10.33480/techno.v19i2.3441

c. Labeling

The labeling process in this study will be done manually using 3 annotators. To find the inter-agreement between the three annotators, the Kappa coefficient is used which is a statistical calculation. The Kappa coefficient value is calculated with an average value of 0.611, where this value indicates a good agreement or has sufficient influence. From the labeling results, the content data on social media labeled Neutral is 27,815, Positive is 5,443 and Negative is 606. The results of the labeling carried out look unbalanced between Positive and Negative Neutrals, where the neutral sentiment value is much higher up to 82.07%, while for the positive it is 16.13% and for the negative it is 1.80%. For this reason, in this study, data sampling will be carried out using the Random UnderSample (RUS) technique and the Synthetic Minority Over-sampling Technique (SMOTE). Random under sample is used to balance the abundant class, while SMOTE is used to increase the missing data. The use of these two data sampling is very important to eliminate over fitting results so that the accuracy value obtained is maximized. After sampling the data, the results obtained for a neutral value of 27815, positive of 18046 and negative of 10866.

d. Word2Vec

Word2Vec is a model name word vector representation made by Google that can represent the meaning of a word and can measure several vectors as a comparison(Sung et al. 2020). In this research, text data that has been pre-processed is processed using the Embedding Layer on Word2Vec with the English and Indonesian language corpus obtained from the nlpl vector site repository. After processing and obtaining the vector of the sentence, the data will be processed using LSTM so that the score and accuracy can be searched(El-diraby, Shalaby, and Hosseini 2019; Rojas-Barahona 2016).

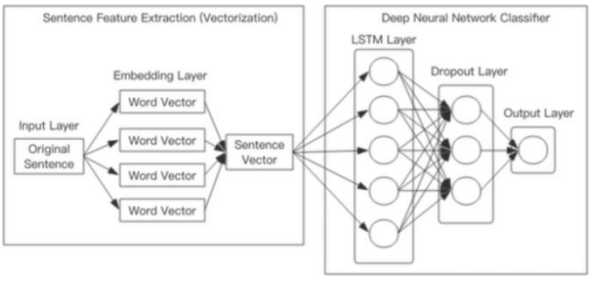

Figure.3 Word2Vec Architecture with LSTM Model

Word2Vec generates a vector space obtained from the corpus, which consists of words that are similar in the corpus and adjacent to each other in the Word2Vec space(Nawangsari et al. 2019). The main principle of this method is to study the laws of dimensional vectors, where Word2Vec can predict words based on their context using one of 2 different neural models namely CBOW and Skip-Gram. Continuous bag of words (CBOW) predicts the current word based on its context. In the CBOW process, three layers are used. Input layer according to the context. The hidden layer corresponds to the projection of each word from the input layer into the weight matrix which is projected to the third layer which is the output layer. The final step of this model is a comparison between the output and the word itself to correct its representation based on the back-propagation of the error gradient(Naili et al. 2017). The process in CBOW can be described by the following equation :

$$\frac{1}{V}\sum_{t=1}^{V} = \log p(\, m_1 | m_{t-\frac{c}{2}} \dots m_{t-\frac{c}{2}}) \quad \text{.....(1)}$$

Skip-Gram is the opposite of CBOW where Skip-Gram looks for context prediction given a word, not word prediction given context like CBOW. The process of Skip-Gram can be described by the following equation:

$$\frac{1}{V}\sum_{t=1}^{V} = \sum_{j=t-c, j\neq t}^{t+c} \log p(\, m_j | m_t) \quad \text{.....(2)}$$

where V corresponds to vocabulary size, c corresponds to the window size of each word.

e. GloVe

The GloVe algorithm is an extension or extension of the word2vec method for efficiency in word vector learning that works by capturing global statistics and local statistics from a corpus. GloVe is an unsupervised learning method to get vector representations for words. Training or training is carried out on globally aggregated words collected from the corpus and the resulting representation displays the linear structure of the word vector space. In this research text data that has been pre-processed is processed using Word Embedding on GloVe. Then the data will be processed using LSTM so that the score and accuracy can be searched(Rojas-Barahona 2016; Song, Park, and Shin 2019).

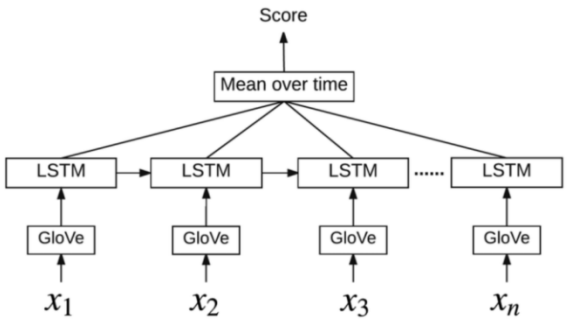

Figure.4 GloVe Architecture with LSTM model

GloVe is an unsupervised learning method for obtaining vector representations for words. Training or training is carried out on global aggregated words collected from the corpus and the resulting representation displays the linear structure of the word vector space. Due to the high quality of textual features, GloVe embedding has been widely used for text mining or natural language processing tasks(Sakketou and Ampazis 2020). In his research, (Levy, Goldberg, and Dagan 2015) stated that there are 2 stages in the GloVe process. The first is the matrix construction of occurrence of X from the training corpus where : is the frequency of the word i that occurs together with the word j. $X_{ij} = \sum_K^V X_{ik}$ is the number of occurrences of the word I in the corpus. The second step is to factor X to get a vector, with the following equation :

$$F(W_i - W_j, W_k) = \frac{P_{ik}}{P_{jk}} \quad \text{.....(3)}$$

Where W_i, W_j and W_k are three vector words, $P_{ik} = X_{ik}/X_i$ is the probability of the word k appearing in the context of the word i, w is the word vector and W_k is the context word vector.

f. AdaComp

AdaComp is a compression technique that can be used for word embedding(Li et al. 2021; May and Labs 2008). AdaComp is based on localized gradient residual selection and automatically adjusts compression level depending on local activity(Kim et al. 2020). AdaComp adaptively adjusts compression ratios across various mini-batches, epochs, network layers, and bins. This characteristic provides automatic adjustment of the compression ratio, resulting in strong model convergence. This technique works in several stages, the first is to find the maximum residual value in each bin. Furthermore, quantitation of the compressed residual vector is carried out to increase the overall compression rate. AdaComp is applied to each layer separately. In this research, text data that has been pre-processed is processed using the Embedding Layer on Word2Vec with the Indonesian language corpus obtained from the nlpl vector site repository. The data that has been processed will be compressed and decompressed so that it can be continued for learning and calculating the accuracy value.

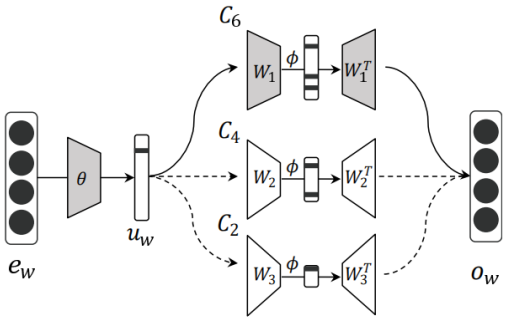

Figure.5 Adaptive Compression with LSTM Model

a. Evaluation Method

The evaluation method used in this study is the Confusion Matrix which aims to obtain the accuracy, precision, recall and F-Score valuesof the models that have been made. Accuracy is the most commonly used evaluation method. Accuracy obtained is the percentage of identified data compared to the sum of all data (Sindhu and Vadivu 2020). The accuracy obtained is the percentage of identified data compared to the sum of all data, to determine the accuracy the following formula is used.

$$Accuration = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{.....(4)}$$

$$\text{Pr}ecission = \frac{TP}{TP + FP} \quad \text{.....(5)}$$

$$\text{Re}call = \frac{TP}{TP + FN} \quad \text{.....(6)}$$

$$F - Score = \frac{2(\text{Pr}ecission * \text{Re}call)}{(\text{Pr}ecission + \text{Re}call)} \quad \text{.....(7)}$$

RESULTS AND DISCUSSION

To determine which method of Word2Vec and GloVe with adaptive compression (AdaComp) is better to improve the accuracy, we conduct research with 4 different models. The model consist of Word2vec model architecture,GloVe model architecture, Word2Vec with AdaComp model architecture and GloVe with AdaComp model architecture. This experimental study employs 3

Vol. 19, No. 2 September 2022 | DOI https://doi.org/10.33480/techno.v19i2.3441

social media dataset in the English and Indonesian language as data in the amount of 33,864 data combination. The details of the data distribution after sampling data were 18046 positive labelled review data, 10886 negative labelled data and 27815 neutral data as depicted in Figure 6. This amount is felt to be sufficient to perform word embedding training in sentiment analysis for pharmaceutical company because it has covered various aspects that are usually assessed by customer, such as service, satisfaction, improvement, benefits, promo, and efficacy.

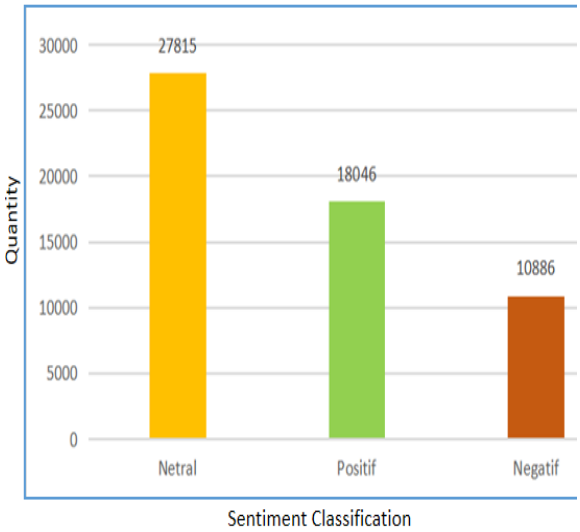

Figure.6 Datased used after sampling data

The initial experiments we use 4 models were carried out using a random under sampling technique and resulted in 100% accuracy for both the GloVe or GloVe models combined with AdaComp. In the Word2Vec model the accuracy obtained is 79% while the use of AdaComp increases the accuracy by 1% so that it is obtained 80%. The use of Random Under Sampling reduces the over fitting, but it is not optimal, so the authors add the Synthetic Minority Over-sampling Technique(SMOTE).

Random Under Sampling (RUS) works by randomly selecting samples from a large group and removing them from the data set, this will continue to be done randomly until a balanced distribution is achieved. Whereas SMOTE works by selecting the instance closest to the feature space, SMOTE first selects a minority class instance at random and finds its k nearest minority class neighbors. A synthetic example is then created by choosing one of the k nearest neighbors at random and connecting to form a line segment in the feature space. The synthetic instance is generated as a convex combination of the two selected instances a and b.

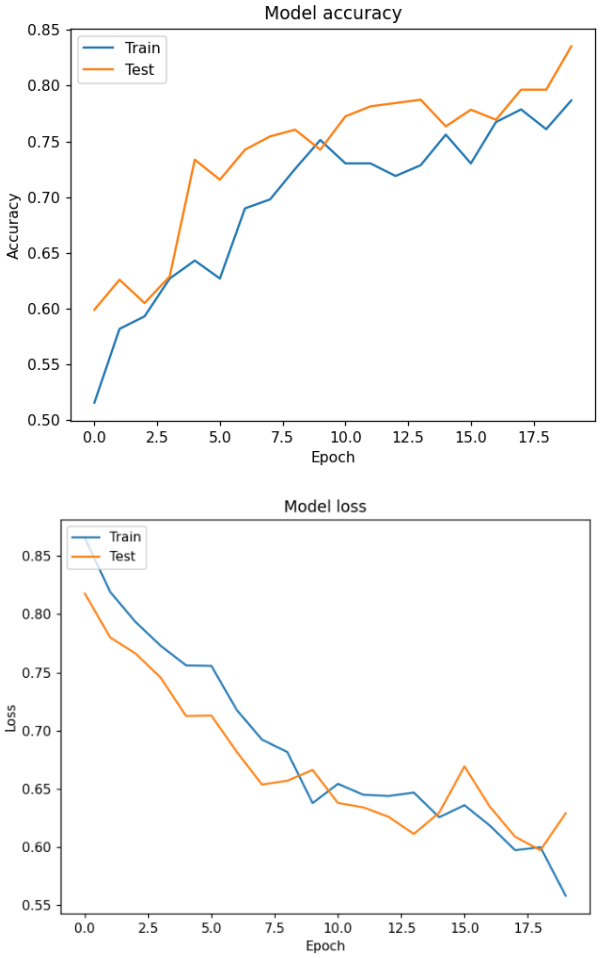

Figure.7 Accuracy model and loss model with RUS data sampling

The use of a smote has a huge impact on the resulting data. In the figure.8 we can see that the data that was previously overfitted becomes good model with the addition of a smote data sampling. As this is the best model as we get.

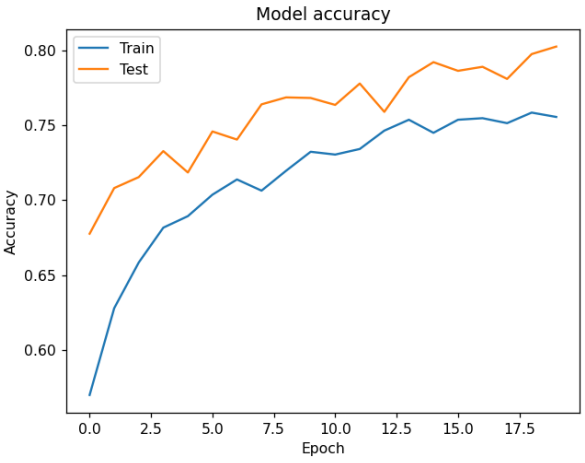

Figure.8 Model Accuracy using RUS and SMOTE data sampling

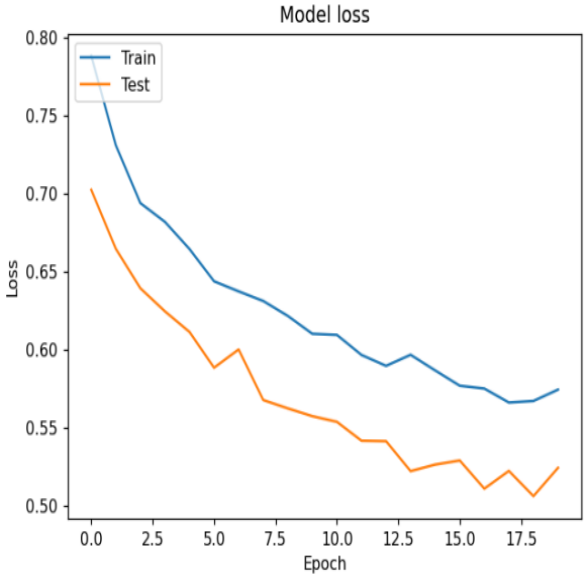

Figure.9 Model Loss using RUS and SMOTE data sampling

The accuracy obtained in both the GloVe and Word2Vec models increasing with the use of AdaComp compression. Adacomp works for each word directly learns to select its code length in an end-to-end manner by applying the Gumbel-softmax tricks. After selecting the code length, each word learns discrete codes through a neural network with a binary constraint.The accuracy results obtained on GloVe is 99%, while with the use of adacom the accuracy value increases from 1% to 100%. Then for the Word2Vec model the accuracy obtained is 80% and there is an increase of 1% with the use of adacomp compression. This summary of this result is shown in table.1

Table 1. Result

Model	RUS		RUS+SMOTE	
	Acc	Score	Acc	Score
GloVe	100	1	99	3
GloVe+AdaComp	100	1	100	1
Word2Vec	79	57	77	50
Word2Vec+AdaComp	80	56	81	49

With the use of adacomp compression the accuracy value generated on Word2Vec with RUS sampling on the evaluation shows the accuracy value increased from 77% to 81% with a True Negative value of 101, False Positive of 18, False Negative of 28 and True Positive of 92 as shown in figure 10.

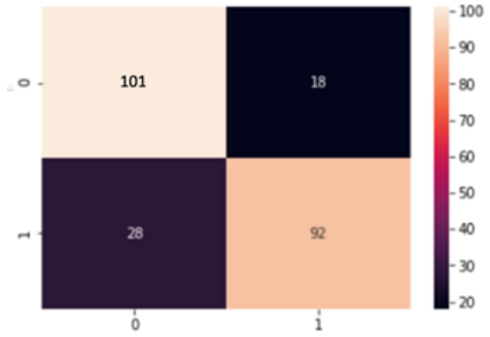

Figure.10 Evaluation for word2vec using adacomp model

```
8/8 [==============================] - 1s 13ms/step
              precision    recall  f1-score   support

           0       0.91      0.61      0.73       120
           1       0.70      0.94      0.81       119

    accuracy                           0.77       239
   macro avg       0.81      0.77      0.77       239
weighted avg       0.81      0.77      0.77       239
```

Figure.11 Evaluation Result for word2vec without adaptive compresion (AdaComp) with RUS and SMOTE

```
8/8 [==============================] - 0s 9ms/step
              precision    recall  f1-score   support

           0       0.78      0.85      0.81       119
           1       0.84      0.77      0.80       120

    accuracy                           0.81       239
   macro avg       0.81      0.81      0.81       239
weighted avg       0.81      0.81      0.81       239
```

Figure.11 Evaluation Result for word2vec with adaptive compresion (AdaComp) with RUS and SMOTE

From the results of the research above, the best model that can increase the highest accuracy is the word2vec model with the use of adacomp and the RUS and SMOTE sampling methods with an increase in accuracy from 77% to 81%. The use of the compression technique on adacomp has been shown to significantly improve the accuracy of the research model used, either using the RUS sampling technique or the combination of RUS and SMOTE. This is possible because the compression technique on adacomp using the Gumbel-Softmax distribution improves the workings of generic word embedding so that for each corpus used approaches the discrete data sampling process, then trained using backpropagation and choosing the length code each word adaptively.

b. Comparison with related works in compression word embedding for sentiment analysis

To evaluate the performance of our models (AdaComp), we conduct a comparison with related works in table 2.In fact, for each language (English and Indonesian language), we conduct an evaluation on the same corpus. For the english language (Table 4), we can state that the use of adacomp give more accuracy. This claim explained by the fact that adding external resource that improves better adaptive compression. In this study we use pretrained word embedding for english version. For the Indonesian language (Table 3) we only compared our work with the work of Nawangsari et al[18]. As shown in Table 3, we notice that the use of adacomp for word2vec in Indonesian embedding is less effective to improve the accuracy.

Table 2. Comparison with existent English topic sentiment analysis

Approach	English word Embedding	Accuracy Score
Generic Word Embedding	Word2Vec	71%
	GloVe	80.23%
Generic Word Embedding with AdaComp	Word2Vec	81%
	GloVe	100%

Table 3. Comparison with existent Indonesian topic sentiment analysis

Approach	Indonesian Word Embedding	Accuracy Score
Generic Word Embedding	Word2Vec	85.96%
	GloVe	77%
Generic Word Embedding with AdaComp	Word2Vec	80%
	GloVe	99%

Based on this evaluation, we can conclude that using adacomp are much way better than using general word embedding without compression for both Indonesian and English languages. This can be explained by the fact that adding external knowledge enhances the quality of adaptive compression. Furthermore, we notice that prediction-based embedding methods improve adaptive compression.

CONCLUSION

This paper focused on adacomp usage with sentiment-specific word embedding that used for pharmaceutical company sentiment analysis. Future studies should investigate the effectiveness of the proposed adaptive compression method for other word embedding model, especially for traditional word embedding sentiment analysis.

The use of compression with the AdaComp model is considered to significantly improve accuracy. This model in proven in the increase of accuracy in the research that the author did, both on imbalanced datasets and balanced datasets. In the imbalanced dataset, there is an increase in accuracy of 2%, 4% and 5% for each dataset used. However, in the best results with a balanced dataset, there is an increase in accuracy in the word2vec model with adacomp using either the RUS sampling method or a combination of RUS and SMOTE.

References

Alam, Muhammad, Fazeel Abid, Cong Guangpei, and L. V. Yunrong. 2020. "Social Media Sentiment Analysis through Parallel Dilated Convolutional Neural Network for Smart City Applications." *Computer Communications* 154:129–37. doi: 10.1016/j.comcom.2020.02.044.

Bagheri, Ayoub, Mohamad Saraee, and Franciska De Jong. 2013. "Knowledge-Based Systems Care More about Customers : Unsupervised Domain-Independent Aspect Detection for Sentiment Analysis of Customer Reviews." *KNOWLEDGE-BASED SYSTEMS* (August). doi: 10.1016/j.knosys.2013.08.011.

Chowdhury, Priyabrata, Sanjoy Kumar Paul, Shahriar Kaisar, and Md. Abdul Moktadir. 2021. "COVID-19 Pandemic Related Supply Chain Studies: A Systematic Review." *Transportation Research Part E: Logistics and Transportation Review* 148(August 2020):102271. doi: 10.1016/j.tre.2021.102271.

De, Rahul, Neena Pandey, and Abhipsa Pal. 2020. "International Journal of Information Management Impact of Digital Surge during Covid-19 Pandemic : A Viewpoint on Research and Practice." *International Journal of Information Management* (June):102171. doi: 10.1016/j.ijinfomgt.2020.102171.

El-diraby, Tamer, Amer Shalaby, and Moein Hosseini. 2019. "Linking Social , Semantic and Sentiment Analyses to Support Modeling Transit Customers ' Satisfaction : Towards Formal Study of Opinion Dynamics." *Sustainable Cities and Society* 49(March):101578. doi: 10.1016/j.scs.2019.101578.

Fabris, M. A., D. Marengo, C. Longobardi, and M. Settanni. 2020. "Investigating the Links between Fear of Missing out, Social Media Addiction, and Emotional Symptoms in Adolescence: The Role of Stress Associated with Neglect and Negative Reactions on Social

Media." *Addictive Behaviors* 106:106364. doi: 10.1016/j.addbeh.2020.106364.

Jiao, Peiran, André Veiga, and Ansgar Walther. 2020. "Journal of Economic Behavior and Organization Social Media , News Media and the Stock Market R." *Journal of Economic Behavior and Organization* 176:63–90. doi: 10.1016/j.jebo.2020.03.002.

Kemp, Simon. 2020. "DIGITAL 2020: 3.8 BILLION PEOPLE USE SOCIAL MEDIA." 2507(February):1–9.

Kim, Yeachan, Kang-Min Kim, and SangKeun Lee. 2020. "Adaptive Compression of Word Embeddings." 3950–59. doi: 10.18653/v1/2020.acl-main.364.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. "Improving Distributional Similarity with Lessons Learned from Word Embeddings." *Transactions of the Association for Computational Linguistics* 3:211–25. doi: 10.1162/tacl_a_00134.

Li, Shuangyin, Rong Pan, Haoyu Luo, Xiao Liu, and Gansen Zhao. 2021. "Knowledge-Based Systems Adaptive Cross-Contextual Word Embedding for Word Polysemy with Unsupervised Topic Modeling." *Knowledge-Based Systems* 218:106827. doi: 10.1016/j.knosys.2021.106827.

Liu, Xia, Hyunju Shin, and Alvin C. Burns. 2019. "Examining the Impact of Luxury Brand ' s Social Media Marketing on Customer Engagement : Using Big Data Analytics and Natural Language Processing." *Journal of Business Research* (April):1–12. doi: 10.1016/j.jbusres.2019.04.042.

May, C. L., and Red Cat Labs. 2008. "Compressing Word Embeddings."

Muhammad, Putra Fissabil, Retno Kusumaningrum, and Adi Wibowo. 2021. "ScienceDirect ScienceDirect Sentiment Analysis Using Word2vec And Long Short-Term Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews Memory (LSTM) For Indonesian Hotel Reviews." *Procedia Computer Science* 179(2020):728–35. doi: 10.1016/j.procs.2021.01.061.

Naili, Marwa, Anja Habacha, Ben Ghezala, and Ben Ghezala. 2017. "ScienceDirect Word Embedding Methods in Topic Comparative Word Comparative Study Study of of Segmentation Word Embedding Embedding Methods Methods in in Topic Topic Comparative Study of Segmentation Word Embedding Methods in Topic Segmentation." *Procedia Computer Science* 112:340–49. doi: 10.1016/j.procs.2017.08.009.

Nawangsari, Rizka Putri, Retno Kusumaningrum, and Adi Wibowo. 2019. "Word2vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study." *Procedia Computer Science* 157:360–66. doi: 10.1016/j.procs.2019.08.178.

Prasad, Poonam J., and G. L. Bodhe. 2012. "Trends in Laboratory Information Management System." *Chemometrics and Intelligent Laboratory Systems* 118:187–92. doi: 10.1016/j.chemolab.2012.07.001.

Rasool, Aaleya, Farooq Ahmad Shah, and Jamid Ul Islam. 2020. "Of." *Current Opinion in Psychology*. doi: 10.1016/j.copsyc.2020.05.003.

Robiady, Nurlita Devian, Nila Armelia Windasari, and Arfenia Nita. 2020. "Customer Engagement in Online Social Crowdfunding : The in Fl Uence of Storytelling Technique on Donation Performance." *International Journal of Research in Marketing* (xxxx):1–9. doi: 10.1016/j.ijresmar.2020.03.001.

Rojas-Barahona, Lina Maria. 2016. "Deep Learning for Sentiment Analysis." *Language and Linguistics Compass* 10(12):701–19. doi: 10.1111/lnc3.12228.

Sakketou, Flora, and Nicholas Ampazis. 2020. "A Constrained Optimization Algorithm for Learning GloVe Embeddings with Semantic Lexicons." *Knowledge-Based Systems* 195:105628. doi: 10.1016/j.knosys.2020.105628.

Sarma, Prathusha K., Yingyu Liang, and William A. Sethares. 2018. "Domain Adapted Word Embeddings for Improved Sentiment Classification." 51–59.

Sindhu, C., and G. Vadivu. 2020. "AUGMENTED KNOWLEDGE SEQUENCE-ATTENTION FINE GRAINED SENTIMENT POLARITY CLASSIFICATION USING AUGMENTED KNOWLEDGE SEQUENCE-ATTENTION MECHANISM." *Microprocessors and Microsystems* 103365. doi: 10.1016/j.micpro.2020.103365.

Song, Minchae, Hyunjung Park, and Kyung shik Shin. 2019. "Attention-Based Long Short-Term Memory Network Using Sentiment Lexicon Embedding for Aspect-Level Sentiment Analysis in Korean." *Information Processing and Management* 56(3):637–53. doi: 10.1016/j.ipm.2018.12.005.

Størdal, Ståle, Gudbrand Lien, Ørjan Mydland, and Erik Haugom. 2021. "Effects of Strong and Weak Non-Pharmaceutical Interventions on Stock Market Returns : A Comparative Analysis of Norway and Sweden during the Initial Phase of the COVID-19 Pandemic." *Economic Analysis and Policy* 70:341–50. doi: 10.1016/j.eap.2021.03.009.

Sung, Yunsick, Sejun Jang, Young Sik Jeong, and Jong Hyuk (James J.). Park. 2020. "Malware Classification Algorithm Using Advanced Word2vec-Based Bi-LSTM for Ground Control Stations." *Computer Communications* 153(December 2019):342–48. doi: 10.1016/j.comcom.2020.02.005.