Extraction of the Aspect of Numerical Entities for Financial News Mining in Bahasa Indonesia

Riki Akbar
200475295
MSc Big Data Science, QMUL
r.akbar@se21.qmul.ac.uk
Project Supervisor: Dr. Julia Ive
j.ive@qmul.ac.uk

***Abstract*—Unlike non-numerical entities, the number of research concentrating on information extraction of numerical entities is still limited despite a massive increase of data-driven approaches touching a wide range of sectors, including the finance domain. Many previous studies emphasised the use of pre-defined keywords to extract attributes and relations with which the numerical entities involve. This approach is less adaptive in capturing the context in which numbers and quantities appear, considering the flexible nature of numerical entities that allows them to fit in various contextual settings. This project proposes two extraction systems, namely NEAE-1 and NEAE-2, that adopt a high-dimensional multilabel classification approach and IndoBERT, a pre-trained BERT language model developed explicitly for Bahasa Indonesia, to tackle the extraction of numerical entity aspects in Indonesian financial news. The evaluation incorporates comprehensive observation of how the proposed system tackles numerous identified challenges: (1) extracting implicit aspects and change attributes and (2) extracting all relevant aspects for numerical entities representing multiple different aspects. The result shows that the proposed extraction systems outperform the baselines with a gain of up to 41.4% in the F1 score. Furthermore, all systems incorporating the IndoBERT achieve a higher F1 score than the systems incorporating the Multilingual BERT with a reported F1 score gain of up to 10.5%, indicating a leveraged performance resulting from a better contextual representation produced by IndoBERT.**

***Index Terms*—numerical entity, aspect, indobert**

I. Introduction

The increasing data-driven approaches to studying financial activities have turned financial texts into an essential source for accessing all related information. Financial news, in particular, has been rapidly emerging as one of the primary sources for obtaining the latest updates on market trends and corporate actions. To acquire this information from financial news, one must observe numerical indicators and determine the relevant attributes or features by learning the context in which the numerical indicators appear. Nevertheless, the financial news is unstructured, posing a significant challenge in extracting the numerical data and transforming them into fruitful information (PwC, 2018).

Most research on information extraction concentrates on non-numerical entities (Madaan et al., 2016), leading to a situation where similar research on numerical entities is limited. The number of research focusing on numerical entities for languages other than English is even lower. For example, there is only a handful of research focusing on relation extraction in Bahasa Indonesia (Indonesian) despite being spoken by well over 200 million people (Adriani and Manurung, 2012). Moreover, the research on extracting the numerical entity aspect, i.e., the word span numerical entity represents, was never previously attempted on any Indonesian corpus.

The main challenge of extracting attributes and relations for numerical entities emerges from their characteristics. Unlike non-numerical entities, numerical entities can fit various contextual settings (Madaan et al., 2016). Take the following sentence as an example: "*Rasio payout pada tahun fiskal 2022 meningkat 15% dibandingkan tahun lalu, sementara margin keuntungan bersih bertahan pada angka 59%*" (*The payout ratio in fiscal 2022 increased 15% compared to last year, while the net profit margin remains at 59%*). The first numerical entity (15%) refers to the payout ratio, while the second (59%) represents the net profit margin. The numerical value of both numerical entities also points out two different types of *change* attribute. The value of the first numerical entity signifies the rate by which the payout ratio increased from last year. In contrast, the value of the second numerical entity specifies the net profit margin achieved for the current fiscal year instead of representing the value change. From this example, it is visible that the flexible nature of numerical entities leads to the requirement of precise contextual representation to distinguish and extract the numerical entity aspect accurately and in a consistent manner.

This project aims to study and tackle the aforementioned challenge of extracting numerical entity aspects from Indonesian financial news. This work proposes two supervised systems incorporating IndoBERT (Koto et al., 2020b) to produce a contextual representation of the sentences in which the numerical entities appear. The first system, **Numerical Entity Aspect Extractor 1 (NEAE-1)**, extracts the main aspects, namely *aspect1* and *aspect2*, and numerical change attributes using a high-dimensional multilabel classification approach and comprises three single-label models which trained for different objectives: (1) extracting the primary aspect (i.e., *aspect1*); (2) extracting the second aspect (i.e., *aspect2*) with which a numerical entity is likely to be associated; and (3) extracting the *change* attribute which a numerical entity

express. On the other hand, the second system, **Numerical Entity Aspect Extractor 2 (NEAE-2)**, is developed based on a similar intuition to the NEAE-1 with a slight difference in terms of taking into account the likelihood of a numerical entity to represent *aspect2* before assigning it with the predicted *aspect2*. Despite the difference, both systems fine-tune the underlying BERT model, i.e., IndoBERT.

The **main contributions** of this project are: (1) Two supervised methodologies to extract the aspect and change attribute of numerical entities on Indonesian corpus, mainly on financial news, and (2) Methodology for annotating the numerical entity samples under the high-dimensional multilabel setting, and (3) The provision of the labelled dataset for use in numerical entity aspect extraction tasks following a multilabel ground-truth tuple of (*aspect1*, *aspect2*, *change*)

II. Related Work

The work on the relation and attribute extraction tasks primarily focused on non-numerical entities, and the distant supervision approach was widely used in many relation extraction systems (Mintz et al., 2009, Riedel et al., 2010). This approach tackled some of the problems both supervised and unsupervised approaches suffer, particularly in relying on the availability of expensive labelled training data and the difficulty in mapping the relations required for specific domains. This approach utilises a sizeable semantic database, namely Freebase (Bollacker et al., 2008), to extract the relation of a pair of entities in a sentence, provided that the pair of entities are involved in a known Freebase relation.

One of the first works on information extraction involving numerical entities is the work of Davidov and Rappoport (2010), where a pattern-based framework was introduced to approximate numerical object properties and improve precision. The framework comprises three stages for listing similar objects, extracting attribute values of these objects using a set of lexical patterns, and feeding the information into approximation.

Roy et al. (2015) proposed a numerical entity representation, Quantity-Value Representation (QVR), laying the foundation for quantitative reasoning in textual inference tasks. With QVR, a numerical entity is represented as a tuple of three components: value, units, and change. The latter explicitly signifies how a numerical entity value is relative to another in a given sentence and offers an opportunity to extract contextual information that numerical entities represent.

The QVR was later adopted to tackle the extraction of numerical relations using rule-based and probabilistic graphical approaches (Madaan et al., 2016). A set of indicative words such as 'surge', 'grow', and 'increase' is specified to extract the change aspect of any numerical entities that appear in a sentence. Another list of one to four keywords per relation is also provided as features to determine the most likely relation at which the numerical entities participate.

P R et al. (2015) also adopted similar approach by utilising a set of domain-specific keywords to extract the numerical attributes and values from a certain type of clinical text, i.e., discharge summary records. Two independent models are proposed for the purpose of establishing a relationship between those numerical attributes and values. The first model, a CRF-based model, takes multiple lexical features as the input, while the second model implements a Support Vector Machine (SVM) and utilises syntactical features to fit the relation that maps the attribute and the value of a numerical entity.

A recent work of Tanwar et al. (2022) incorporates a BERT-based model, i.e., ClinicalBERT (Huang et al., 2019), with external knowledge to address the challenges of phenotypes extraction from clinical texts, particularly in which numerical reasoning is required. The proposed workflow, namely the Numerical Reasoning (NR) Model, takes numeric entities from the external knowledge and the extracted lexical candidates from the text to produce contextualised embeddings. The generated contextualised embeddings are used in deciding which numeric entity to be assigned to the extracted lexical candidates and determining the corresponding phenotypes.

III. Background

A. Numerical Entities in Finance Domain

The challenges in dealing with numbers and quantities in texts mainly originated from their unique characteristics. In contrast to dealing with non-numerical entities, capturing any regularities involving numerical entities is challenging as they can participate in virtually any relation with other entities (Madaan et al., 2016). In financial texts, numerical entities can represent any aspect of information, from the stock price to the number of unemployment in a country, in either explicit or implicit settings. For instance, in the following sentence, "*Putra Rajawali Kencana memperoleh laba kotor sebesar Rp 17,34 miliar atau 32,87% lebih besar dari periode yang sama tahun sebelumnya Rp 13,05 miliar.*" (*Putra Rajawali Kencana recorded a gross profit of Rp 17,34 billion or increased 32,87% from the gross profit recorded at the same period last year, Rp 13,05 billion*), the quantity **Rp 13,05 miliar** not only represent the gross profit from last year but also serves as a context from which the current year gross profit increased.

A numerical entity may represent more than one aspect, depending on the context of the sentence. A numerical entity aspect may serve as a hypernym to the others. For instance, in the sentence: "*Berdasarkan laporan keuangannya, MAPI melaporkan kenaikan laba bersih sebesar 16,04% dari Rp 16,3 Triliun menjadi 18,92 Triliun.*"(*According to its financial statement, MAPI reports a net income increase of 16.04% from Rp 16.3 Trillion to 18.92 Trillion.*) the quantity **18,92 Triliun** can represent two aspects: the net income and income. Although it is more precise to say that the numerical entity (i.e., 18,92 Triliun) refers to the net income, one may argue that it can also refer to the company income in general as the net income is essentially the income itself.

Alternatively, a numerical entity aspect may appear independently and complements the other extracted aspect. For example, the quantity **Rp 7.800/share** in the following sentence: "*Harga pembukaan BMRI sama dengan harga penutupan sehari sebelumnya, tepatnya pada harga Rp 7.800/saham.*"

(*The opening price of BMRI is Rp 7.800/share, the same as the closing price the day before.*) refers to both opening and closing price as the opening price of today trading is equal to the closing price from the previous trading day. Note that none of both aspects serves as the hypernym to each other.

There are also emerging domain-specific challenges, namely the implicit aspect, frequently occurring when the indicative keywords are not present in a sentence. Consequently, the keyword-based approaches (Madaan et al., 2016, P R et al., 2015) are less effective in dealing with such issues. In addition, the frequent use of financial jargons adopted from different languages might prevent the numerical entity aspects from being extracted.

B. IndoBERT and Multilingual BERT

BERT-based models (Devlin et al., 2019) have been widely adopted and achieve state-of-the-art (SOTA) performance in many downstream tasks. BERT has been proven powerful in providing a contextual representation of texts and outperforms many existing word embeddings that cannot represent the context of any sentence, resulting in the lack of prediction accuracy.

Recent development shows that many pre-trained BERT models are available for solving many downstream tasks. These models offer the advantages of transfer learning without having to train a sophisticated model from scratch. Multilingual BERT, for instance, is a pre-trained multilingual model widely used in tackling NLP tasks for non-English corpus and supports over 104 languages, including Bahasa Indonesia.

As an alternative to Multilingual BERT, (Koto et al., 2020a) introduced IndoBERT, a monolingual pre-trained BERT model developed explicitly for Bahasa Indonesia. IndoBERT is trained on more than 220 million words collected from multiple sources, including Wikipedia and Indonesia Web Corpus (Medved and Suchomel, 2017). IndoBERT quickly emerges as the standard model as it has achieved state-of-the-art (SOTA) performance in many NLP tasks performed on the Indonesian corpus. It even outperforms the previous SOTA, MalayBERT.

In the experimental setup of this work, the proposed systems implement IndoBERT and Multilingual BERT to produce the contextual embeddings of the input sentence containing numerical entities. They are fine-tuned using a convolutional layer with which they achieve better performance than when using the multilayer perceptron (MLP). The evaluation of both BERT models seeks to observe which one is more reliable in providing contextual representation to overcome linguistic challenges encountered in the Indonesian corpus.

IV. Methodologies

A. Datasets and Annotation Procedure

This project aims to tackle the novel task of extracting aspects of numerical entities in multilabel settings. Each numerical entity represents at least two pieces of information: the aspect and *change* attribute. The numerical entity aspect is a particular subject or topic a numerical entity represents, given the context of the sentence in which it appears. In contrast, the change attribute specifies how the value of a numerical entity changes (Roy et al., 2015). For example, in the sentence "*Sekitar 25% dari laba bersih akan dibayarkan sebagai deviden, naik sebesar 15% dari deviden yang dibagikan pada tahun lalu.*" (*It is estimated that 25% of the net profit will be paid as a dividend, increased 15% from the amount of dividend paid last year.*), the quantity **25%** and **15%** represent different types of information. While it is visible that the quantity **25%** refers to a portion of net profit that will be paid to the stakeholders as a dividend, the quantity **15%** signifies the change in the amount of dividend paid to the stakeholders when compared with the payout last year.

Due to the novelty of this task, no annotated dataset is available for use; therefore, the dataset, i.e., financial news, needs to be collected and subsequently annotated using a self-annotating procedure. This project collects and annotates the data without the involvement of any external parties assisting those processes. A semi-automatic annotation scheme is introduced with the regular expression implementation in developing multi-stage annotation rules, equipped with manual verification to ensure the annotation output is accurate and consistent.

In this project, the target aspects are developed by referring to financial and investment dictionaries provided by two portals: **Finansialku.com**[1] and **Bareksa.com**[2]. Further adjustments are implemented to ensure no overlapping terminologies fall under a different aspect.

The publicly available financial news dataset is collected from two leading Indonesian media: **Kontan**[3] and **CNBC Indonesia**[4]. The dataset comprises 10,855 sentences from three specific news channels: investment, financial, and market; each contains at least one numerical entity. These sentences are preprocessed using a standardisation procedure applied to each specific token with the following criteria:

1. **Currency units.** Each currency unit is transformed into a three-letter code following the standard of ISO 4217 Currency Code[5] e.g., *US$* and *US Dollars* is changed to be *USD*.
2. **Date and Time.** Each date and time unit is masked using designated tokens e.g., *12/08/20* is masked to be '[DATE]' and *08:30 WIB* is masked to be '[TIME]'.
3. **Lower case.** All characters in the sentence is transformed into lower case.

As displayed in Figure 1, The annotation consists of several rule-based procedures. First, a numeric token is identified using regular expressions and subsequently combined with any matching numerical units obtained by looking up the numerical unit list acquired from **KBBI**[6] (the official dictionary of the Indonesian language) to extract a numerical entity. In addition,

[1]https://www.finansialku.com/kamus-keuangan
[2]https://www.bareksa.com/kamus
[3]https://www.kontan.co.id/
[4]https://www.cnbcindonesia.com/
[5]https://www.iso.org/iso-4217-currency-codes.html
[6]https://kbbi.kemdikbud.go.id/

a numerical entity type is also determined based on the predefined types specified in Table I. A set of change-related keywords is also used as a reference to extract the change attribute from observing a numerical value that a numerical entity express. At this point, the annotation attempts to link the extracted numerical entity into the most likely aspect it represents. The experiment with an Indonesian dependency parser tool, namely **Morphind** (Larasati et al., 2011), is conducted to perform syntactical analysis to achieve this objective. However, the initial result demonstrates the limitation in establishing dependency for preliminary samples of sentences containing domain-specific terminology. This finding leads to the alternative method, employing naive distance measures to map the numerical entity with the nearest matching keyword from a specific class and assign the corresponding class as the ground-truth aspect. This annotation result is further verified by hand at the final stage to ensure it follows the expected labelling consistently. The manual annotation also involves tagging samples with identified linguistic challenges such as implicit aspects and change attributes, as well as the number of aspects each numerical entity represents, i.e., single or multi aspects.

TABLE I
NUMERICAL ENTITY TYPES

Numerical Entity Type	Description
currency	Any numerical entity representing currency units e.g., USD 98 and GBP 20
percentage	Any numerical entity representing percentage, typically followed by percent sign (%) e.g., 15%
number	Any other numerical entity that does not fall into both types above e.g., 10 *lembar saham* (*10 shares*)
interval_currency	Any numerical entity representing a range or an interval of currency units e.g., USD 20-50
interval_percentage	Any numerical entity representing a range or an interval of percentage e.g., 10-15%
interval_number	Any numerical entity representing a range or an interval of number e.g., 10-20 *kantor cabang* (10-20 *branch offices*)

Instead of annotating individual tokens in a sentence as typically performed in sequence-labelling tasks such as Named Entity Recognition (NER), the annotation procedure applies to each extracted numerical entity as shown in the Figure 1. The annotation starts with a keyword lookup operation to determine whether a keyword from certain aspect classes exists in a preprocessed sentence and subsequently assigns a corresponding class to the numerical entity.

As a result of annotation process, each numerical entity is assigned a tuple of three components, i.e., (*aspect1, aspect2, change*), as its ground truth label, where *aspect1* and *aspect2* denote the primary and secondary numerical entity aspect,

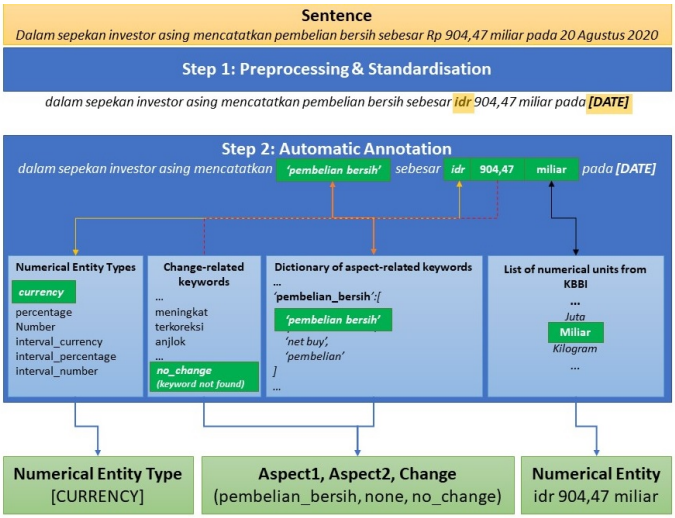

Fig. 1. The annotation procedure of numerical entity. A sentence is preprocessed and standardised by based on three criteria: currency units, date and time masking, and transforming all characters into lower case. In the subsequent step, numerical entity is extracted by performing lookup operation to match any token with specific class keywords. KBBI is also utilised to identify the existence of any numerical units. The result of annotation is a tuple of (*aspect1, aspect2, change*).

and *change* denotes the change attribute. If a numerical entity only represents one aspect, its *aspect2* is assigned with 'none'. Further manual verification is performed by observing random samples to ensure a consistent annotation result.

The distribution of the majority classes from each *aspect1* and *aspect2* are shown in Table II and Table III, respectively. Furthermore, the classes of *change* is shown in Table IV. More details on aspect class distribution are provided in Appendix A.

Note that the numerical entity type plays a pivotal role in establishing a more obvious context in which a numerical entity appears, as relying on the numerical entity value may prevent the embedding, i.e., BERT, from learning and producing an accurate contextual representation. The exploratory data analysis findings confirm this rationale, suggesting a correlation between aspects and certain numerical entity types. For example, the quantities that appear in the form of currency tend to represent the price-related aspect, e.g., the stock and commodity price. On the other hand, quantities that appear as percentages often refer to margins and profits.

B. The Baselines and Proposed Framework

In this project, we select two systems as the baselines. The first baseline model, **Baseline1**, is a multilabel classifier based on a pre-trained T5 model (Raffel et al., 2020) fine-tuned using the Indonesian corpus[7], i.e., **id_liputan6** dataset (Koto et al., 2020a). Baseline1 adopts a sequence-to-sequence approach which views the extraction of numerical entity aspect as a text summarisation problem with a length-constrained

[7]https://huggingface.co/cahya/t5-base-indonesian-summarization-cased

TABLE II
THE DISTRIBUTION OF MAJORITY CLASSES IN *Aspect1*

Class	Description	Number of samples
ihsg	Indonesia Composite Index	2,567
pembelian_bersih	Net buy	849
sp_500	Standard & Poor's 500 Index	691
penjualan_bersih	Net sales	643
dow_jones	Dow Jones Industrial Average	637

TABLE III
THE DISTRIBUTION OF MAJORITY CLASSES IN *Aspect2*

Class	Description	Number of samples
none	a designated class for numerical entity who represents specifically one aspect	15,639
pembelian	Purchase	849
yield	Investment return	740
penjualan	Sales	695
laba	Earning	461

TABLE IV
THE DISTRIBUTION OF CLASSES IN *Change*

Class	Description	Number of samples
decrease_by	indicates a value change **by** which a numerical entity aspect decreases	2557
decrease_from	indicates a value change **from** which a numerical entity aspect decreases	470
decrease_to	indicates a value change **to** which a numerical entity aspect decreases	1647
increase_by	indicates a value change **by** which a numerical entity aspect increases	3842
increase_from	indicates a value change **from** which a numerical entity aspect increases	611
increase_to	indicates a value change **to** which a numerical entity aspect increases	2564
no_change	no indication of any value change, suggesting the numerical value is the value of a numerical entity aspect itself	9048

target sequence. Three T5 models are trained separately with *aspect1*, *aspect2*, and *change* as the targets, respectively.

In contrast, the second baseline model, **Baseline2**, implements BERT (Devlin et al., 2019) to produce a contextual representation of the sentence and the numerical entity type. A convolutional layer is set to fine-tune the underlying BERT model before computing the probability distribution of each aspect using a softmax layer. The top three aspects with the highest score will be selected and inferred as the extracted aspects for the given numerical entity.

This project proposes two numerical entity aspect extraction systems, **NEAE-1** and **NEAE-2**, which adopt a multi-target prediction approach. Both systems comprise several independent models, each trained to predict one target aspect, considering minimal dependencies from one target aspect to the others. This approach follows the preliminary observation finding that a numerical entity aspect is more likely to be determined by the context in a sentence rather than the information of other extracted aspects from other numerical entities.

As shown in the Figure 2, the first system, **NEAE-1**, comprises three models, namely *submodels_mtp_model1*, *submodels_mtp_model2*, and *submodels_mtp_model3*. All of those three models are trained independently to extract each *aspect1*, *aspect2*, and *change*, respectively.

The Figure 3 shows the architecture of the second system, i.e., **NEAE-2**. The NEAE-2 also incorporates *submodels_mtp_model1* and *submodels_mtp_model3* to extract the *aspect1* and the *change* attribute in similar fashion to NEAE-1. However, instead of using a single model to extract the *aspect2*, NEAE-2 trained two models (*submodels_mtp_model2_0* and *submodels_mtp_model2_1*) to determine if a numerical entity represents more than one aspect and assign the most likely *aspect2* when determined so. The intuition of this approach is to minimise the domination of the 'none' class during the learning process by preventing the numerical entities that represent more than one aspect from being assigned with 'none'. As indicated in Table III, more samples represent only one aspect, resulting in the class distribution of *aspect2* being imbalanced.

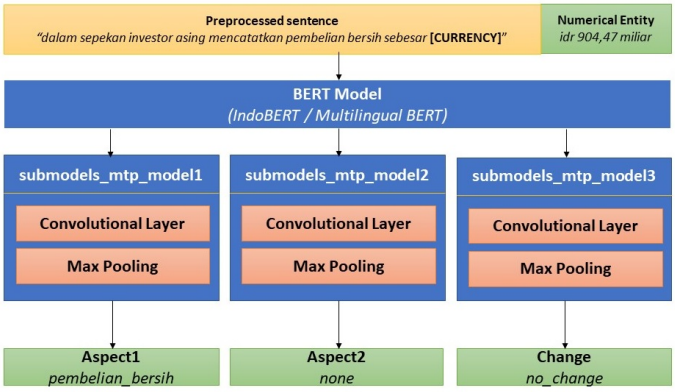

Fig. 2. The first proposed system: NEAE-1. NEAE-1 comprises three models to predict each of *aspect1*, *aspect2*, and *change*.

All of Baseline1, NEAE-1, and NEAE-2 implement BERT as the sentence embedding and a convolutional layer is stacked on top of the BERT as a supervised fine-tuning.

C. Training and Evaluation

There are 20,739 samples of numerical entities divided into 18,520 and 2,219 samples allocated for training and testing, respectively. A small portion of the training dataset, i.e., 10%,

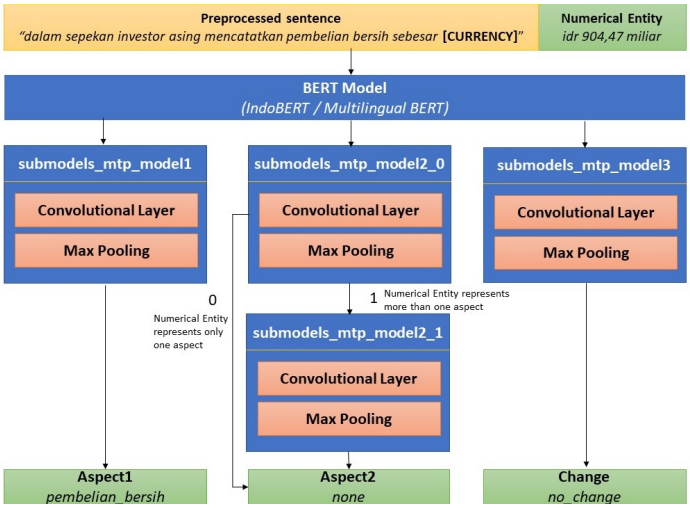

Fig. 3. The first proposed system: NEAE-1. NEAE-1 comprises three models to predict each of *aspect1*, *aspect2*, and *change*.

serves as held-out data. All models are trained independently and involve different experiment setups, including hyperparameter tuning, which affects the learning process significantly. A further investigation of incorporating different BERT models into all systems is conducted to observe the most reliable BERT model in producing contextual representations from the sentences and numerical entity types.

This project seeks to evaluate the performance of all baselines and proposed systems by observing their ability to correctly distinguish positive and negative samples for every class in *aspect1*, *aspect2*, and *change*. The ground truth and the predicted aspects are required to follow the tuple (*aspect1, aspect2, change*), and thus, the inference process assigns *aspect2* for all numerical entities representing only one label with 'none'.

It is visible that the class distribution encountered in *aspect1* and *aspect2* are imbalanced. This condition may lead to a misleading evaluation; hence, selecting appropriate evaluation metrics is essential to prevent the scoring from favouring the majority class (Gowda et al., 2021, Narasimhan et al., 2016). In addition, for multilabel settings, partially-correct predictions must also be considered to capture the actual performance demonstrated by all systems. *Macro averaging* overcomes the stated challenges because it considers each class equally important and, thus, suitable to measure the overall performance of a system. The precision, recall, and f1-score are computed class-wise before aggregating them across all classes in *aspect1*, *aspect2*, and *change*.

The evaluation strategy also incorporates the observation of the ability of the baselines and proposed systems to tackle the identified linguistics issues, including (1) the extraction of implicit aspects and change attributes; and (2) determining all correct aspects for numerical entities representing more than one aspect.

V. Results and Discussion

The evaluation result reported in the Table V shows that the proposed systems outperform baselines in terms of overall performance. The NEAE-2 achieves the best F1-score with gains of 41.4% and 3.3% from Baseline1 and Baseline2, respectively. On the other hand, the Baseline1 records the lowest precision, recall, and F1 scores among all the systems, indicating the inability of the pre-trained T5 model to capture contextual representation for the Indonesian corpus and predict the target aspect as a text that aligns with the defined aspect classes. For example, in the following sentence: "*Nilai transaksi saham di perdagangan kemarin, mencapai Rp 8,27 triliun dengan 13,55 miliar saham ditransaksikan*" (*The total transaction value on the previous trading day reached Rp 8.27 trillion with 13.55 billion shares being transacted*), the Baseline1 fails to extract the *volume_transaksi_saham* (stock transaction volume) as the aspect represented by the quantity **13,55 miliar saham**. Instead, the Baseline1 predicted the *transaksi_saham* (stock transaction) as an aspect represented by the same quantity.

TABLE V
The Overall Performance of the Baselines and Proposed Systems

System	Precision	Recall	F1-score
Baseline1	0.476	0.379	0.402
Baseline2 (mBERT)	0.734	0.717	0.709
Baseline2 (IndoBERT)	0.828	0.777	0.783
NEAE-1 (mBERT)	0.781	0.717	0.716
NEAE-1 (IndoBERT)	0.859	0.812	0.814
NEAE-2 (mBERT)	0.779	0.708	0.711
NEAE-2 (IndoBERT)	**0.86**	**0.815**	**0.816**

The result also shows that the selection of underlying BERT models directly affects the system performance. All systems incorporating IndoBERT as the sentence embedding observe a significant improvement up to 10.5% of the F1-score. The Multilingual IndoBERT implementation leads to slightly low precision, recall, and F1 scores for all systems. This result suggests that IndoBERT produce a better contextual representation of the Indonesian corpus, allowing the systems to extract the numerical entity aspects more precisely.

Another evaluation of the ability of each system to tackle identified linguistics issues is also performed, as reported in the Table VI. All systems are evaluated using their optimal configuration, including the implementation of IndoBERT as an underlying BERT for Baseline2, NEAE-1, and NEAE-2. There are three evaluation objectives: (1) the extraction of implicit aspects and (2) implicit change attributes, and (3) extracting all relevant aspects for numerical entities represent more than one aspect.

In extracting all aspects for multi-aspect numerical entities, the Baseline2 reports the highest F1 score and outperforms the proposed systems. Unlike both proposed systems, which employ separate models to predict each target aspect and

TABLE VI
THE REPORTED F1 SCORE OF THE BASELINES AND PROPOSED SYSTEMS IN TACKLING IDENTIFIED LINGUISTICS ISSUES

System	Implicit aspects	Implicit change	multi-aspect
Baseline1	0.109	0.362	0.111
Baseline2	0.544	**0.469**	**0.468**
NEAE-1	**0.573**	**0.476**	0.35
NEAE-2	**0.573**	**0.476**	0.378

TABLE VII
THE COMPARISON BETWEEN EXTRACTED ASPECT AND THE CORRESPONDING GROUND-TRUTH FOR EACH SYSTEM IN TACKLING SEVERAL IDENTIFIED LINGUISTIC ISSUES

Issue	Input sentence	Numerical Entity	Ground-truth	Baseline1	Baseline2	NEAE-1	NEAE-2
Implicit aspect	*banyak analis mengatakan pemangkasan suku bunga sebesar 25 basis poin bps akan dihargai pasar.*	25 basis poin bps	**(suku bunga the fed)**	(increase by)	(none)	**(suku bunga the fed)**	**(suku bunga the fed)**
Implicit change	*dus, mirae asset mempertahankan asumsi produksi batubara china year dan 2022 masing-masing sebesar 3,49 miliar ton 2,7% secara yoy dan 3,51 miliar ton 0,6% secara yoy.*	2,7%	**(increase by)**	**(increase by)**	(increase from)	**(increase by)**	**(increase by)**
Singe vs multi aspects	*namun, inflow dana asing tetap mengalir deras, terbukti dari net buy asing di seluruh pasar yang mencapai idr 5,37 triliun di sepanjang pekan lalu.*	idr 5,37 triliun	**(dana)**	(pembelian bersih, pembelian)	**(dana)**	(dana pembelian, pembelian)	(dana pembelian, pembelian)

change attribute, Baseline2 enjoys the advantage of having a broader view as a result of implementing softmax to observe the probability distribution of all aspect classes. This strategy enables Baseline2 to determine better whether a particular numerical entity represents one or more than one aspect, given the context of the sentence where it appears. However, the result also indicates that the two-model configuration in NEAE-2 effectively addresses the imbalanced class problem from which the NEAE-1 suffers due to the frequent 'none' class.

Despite the low F1-score in extracting aspects for multi-aspect numerical entities, both proposed systems outperform the baselines in tackling all other identified challenges. The result suggests that based on the F1-score, NEAE-1 and NEAE-2 report the best performance in extracting implicit aspects. Both systems achieve identical performance as a result of employing the *submodels_mtp_model1*, which is trained to extract the *aspect1*. Similarly, the NEAE-1, and NEAE-2 report the same performance in extracting implicit *change* attributes as each system utilises the *submodels_mtp_model3* to extract the change. The comparison between the extracted aspects with the ground-truth for samples with identified issues is shown in Table VII

VI. FUTURE WORK

The limited resource of labelled data will still become a significant challenge in tackling the NLP tasks focusing on numerical entities. One may consider a semi-supervised approach in which the systems are self-trained using a small portion of labelled data. Alternatively, implementing a sequence-to-sequence model, such as the pre-trained T5 model, in generating synthetic sentences may help reduce the expenses of having a large-scale labelled dataset. Thus, the involvement of manual verification could be gradually eliminated, allocating more substantial time for the experiment to shape and improve the proposed systems.

Furthermore, exploring the implementation of Graph Convolutional Networks (Kipf and Welling, 2016) as an alternative

to the convolutional layer to fine-tune the IndoBERT also offers an exciting opportunity to extract more aspects without having to train multi-target models separately. The implementation of Graph Convolutional Networks aims to improve the ability of the extraction systems to extract multiple aspects, replacing the existing intuitive approaches such as the *top-k softmax* incorporated in Baseline2.

Acknowledgments

I would like to express my gratitude to LPDP (Indonesian Endowment Fund for Education) for funding my education and allowing me access to high-quality teaching and research materials, particularly in the field of data science, at the Queen Mary University of London. I would also like to thank my supervisor, Dr Julia Ive, who has been very helpful and supportive throughout my journey of working on this project. Her guidance, feedback, and encouragement have opened my perspective on many aspects of the NLP area and helped me gain further knowledge about my project.

References

Adriani, M. and Manurung, R. (2012). A survey of Bahasa Indonesia NLP research conducted at the University of Indonesia.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge, *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, Association for Computing Machinery, New York, NY, USA, p. 1247–1250.

Davidov, D. and Rappoport, A. (2010). Extraction and approximation of numerical attributes from the web, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Uppsala, Sweden, pp. 1308–1317.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Gowda, T., You, W., Lignos, C. and May, J. (2021). Macro-average: Rare types are important too, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, pp. 1138–1157.
URL: *https://aclanthology.org/2021.naacl-main.90*

Hasudungan, R. J. and Purwarianti, A. (2018). Relation detection for indonesian language using deep neural network - support vector machine, *2018 International Conference on Asian Language Processing (IALP)*, pp. 290–295.

Huang, K., Altosaar, J. and Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission, *arXiv:1904.05342* .

Kamayani, M. and Purwarianti, A. (2011). Dependency parsing for indonesian, *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pp. 1–5.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks, *arXiv preprint arXiv:1609.02907* .

Koto, F., Lau, J. H. and Baldwin, T. (2020). Liputan6: A large-scale Indonesian dataset for text summarization, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, Suzhou, China, pp. 598–608.
URL: *https://aclanthology.org/2020.aacl-main.60*

Koto, F., Rahimi, A., Lau, J. H. and Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP, *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 757–770.

Larasati, S., Kubon, V. and Zeman, D. (2011). Indonesian morphology tool (morphind): Towards an indonesian corpus, Vol. 100, pp. 119–129.

Madaan, A., Mittal, A., Mausam, Ramakrishnan, G. and Sarawagi, S. (2016). Numerical relation extraction with minimal supervision, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, AAAI Press, p. 2764–2771.

Medved, M. and Suchomel, V. (2017). Indonesian web corpus (idwac).

Mintz, M., Bills, S., Snow, R. and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, Suntec, Singapore, pp. 1003–1011.

Narasimhan, H., Pan, W., Kar, P., Protopapas, P. and Ramaswamy, H. G. (2016). Optimizing the multiclass f-measure via biconcave programming, *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1101–1106.

P R, S., Mandhan, S. and Niwa, Y. (2015). Numerical atrribute extraction from clinical texts.

PwC (2018). The data intelligent tax administration: Meeting the challenges of big tax data and analytics.
URL: *https://www.pwc.nl/nl/assets/documents/the-dataintelligent-tax-administrationwhitepaper.pdf*

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* **21**(1).

Rahman, A., Adhiguna, K. and Purwarianti, A. (2017). Ensemble technique utilization for indonesian dependency parser, *PACLIC*.

Riedel, S., Yao, L. and McCallum, A. (2010). Modeling relations and their mentions without labeled text, *ECML/PKDD*.

Roy, S., Vieira, T. and Roth, D. (2015). Reasoning about quantities in natural language, *Transactions of the Association for Computational Linguistics* **3**: 1–13.

Tanwar, A., Zhang, J., Ive, J., Gupta, V. and Guo, Y. (2022). Unsupervised numerical reasoning to extract phenotypes from clinical text by leveraging external knowledge, *arXiv:2204.10202* .

Wan, X., Yang, J. and Marinov, S. (2021). Sentiment correlation in financial news networks and associated market movements, *Sci Rep 11, 3062* .
URL: *https://doi.org/10.1038/s41598-021-82338-6*

Wibawa, A. S. and Purwarianti, A. (2016). Indonesian named-entity recognition for 15 classes using ensemble supervised learning, *Procedia Computer Science* **81**: 221–228. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Zhang, J., Bolanos Trujillo, L., Li, T., Tanwar, A., Freire, G., Yang, X., Ive, J., Gupta, V. and Guo, Y. (2021). Self-supervised detection of contextual synonyms in a multi-class setting: Phenotype annotation use case, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 8754–8769.

Zhao, X., Greenberg, J., An, Y. and Hu, X. T. (2021). Fine-tuning bert model for materials named entity recognition, *2021 IEEE International Conference on Big Data (Big Data)*, pp. 3717–3720.

APPENDIX

A. ASPECT CLASS DISTRIBUTION

Aspect	Description	Number of train samples	Number of test samples
akuisisi	acquisition	19	3
alokasi_dana	fund allocation	38	5
arus_kas	cashflow	17	2
aset	assets	129	15
aset_lancar	current assets	23	3
aset_setara_kas	cash equivalent assets	35	5
aset_tidak_lancar	non-current assets	15	2
asx_200	Australia share market index	201	23
auto_rejection_atas	the maximum limit of stock price increase	25	3
auto_rejection_bawah	the minimum limit of stock price decrease	21	3
belanja	expenditure	15	0
belanja_modal	capital expenditure	62	9
biaya	cost	27	4
biaya_administrasi	administration cost	75	9
biaya_bunga_kredit_investasi	cost of credit and investment interest	16	2
biaya_keuangan	financial cost	64	8
biaya_overhead	overhead cost	40	5
biaya_pendapatan	revenue cost	10	3
biaya_penjualan	sales cost	29	4
biaya_produksi	production cost	14	2
biaya_provisi_appraisal	provision and appraisal cost	8	2
biaya_transaksi	transaction cost	13	2
bopo	operational cost of operating income	67	8
buyback_emas	buyback price of gold commodity	76	9
buyback_saham_obligasi	buyback price of stocks and bond	17	2
cac_40	French stock market index	31	4
cadangan_kas_devisa	Foreign exchange reserves	15	2
casa	current account saving account	17	2
cicilan	installment	9	2
cut_loss	cut loss strategy	9	2
dana	funds	177	20
dana_ipo	funds from IPO	44	5
dana_kelolaan	managed funds	18	2
dana_obligasi_pungutan_talangan	bond fund	8	2
dax	German stock market index	261	30
defisit	deficit	17	2
defisit_anggaran	budget deficit	9	2
denda	fine	17	2
deposito	deposito	17	2

diskon_cashback	discount and cashback	24	3
dividen	dividend	116	14
dividen_saham	dividend stocks	101	12
dividen_tunai	cash dividend	40	5
dow_jones	Dow Jones industrial average	573	64
dpk	third-party funds	104	12
ebitda	earnings before interest, taxes, depreciation and amortization	53	7
fasilitas_kredit	credit facility	20	3
fee	fees	12	2
ftse_100	Financial Stock Exchange 100 index	65	8
gagal_bayar	default	16	2
gaji_tunjangan_kompensasi	income and compensation	99	12
giro	giro	13	2
gross_profit_margin	gross profit margin	42	5
hadiah_bantuan_stimulus	aid and stimulus	21	3
hak_suara	voting rights	9	2
hang_seng	Hongkong stock market index	280	32
harga	price	37	5
harga_batubara	coal price	77	9
harga_bitcoin	bitcoin price	170	20
harga_cpo	cpo price	169	19
harga_emas	gold price	167	19
harga_ethereum	ethereum price	146	17
harga_komoditas_lain	other commodity price	34	4
harga_kripto_lain	other cryptocurrency price	146	17
harga_minyak	oil price	257	29
harga_nikel	nickel pricenick	45	6
harga_penawaran	offering price	36	5
harga_rumah_properti	property price	10	2
harga_saham	stock price	336	38
harga_saham_ipo	IPO stock price	28	4
harga_saham_pembukaan	stock price at the opening of trading day	22	3
harga_saham_penutupan	stock price at the closing of trading day	24	3
harga_saham_terendah	lowest stock price	25	3
harga_saham_tertinggi	highest stock price	24	3
harga_timah	lead price	29	4
ihsg	Indonesia composite index	2310	257
indeks	index	34	4
indeks_harga_konsumen	consumer price index	10	2
inflasi	inflation	74	9
jumlah_debitur	the number of debtor	20	3
jumlah_ekspor	the number of export	62	8

jumlah_impor	the number of import	19	3
jumlah_investor	the number of investor	30	4
jumlah_kantor	the number of office	10	2
jumlah_karyawan	the number of employees	14	2
jumlah_nasabah	the number of customers	58	7
jumlah_outlet_agen	the number of agent outlets	16	2
jumlah_pekerjaan	the number of jobs	9	2
jumlah_pelanggan_pengguna	the number of users	13	2
jumlah_rekening	the number of account	8	2
jumlah_rumah_properti	the number of properties	9	2
jumlah_umkm	the number of SMEs	11	2
kapasitas	capacity	26	4
kapasitas_produksi	production capacity	26	4
kapitalisasi_pasar	market capitalisation	145	17
kekayaan	wealth	92	11
kepemilikan	ownership	166	19
kepemilikan_saham	shares	14	2
kinerja	performance	8	2
kinerja_reksadana	mutual funds performance	10	2
kinerja_sektor	sectoral performance	34	4
klaim	claim	34	4
klaim_tunjangan_pengangguran	claim of unemployment benefits	24	3
komisi	commission	13	2
kontribusi	contribution	14	2
kontribusi_ekspor	export contribution	43	5
kontribusi_pendapatan	revenue contribution	53	6
kontribusi_penjualan	sales contribution	26	5
kospi	Korea composite stock price index	230	26
kredit_macet	bad credits	69	8
kupon_obligasi	bond coupon	10	2
kurs	exchange rate	144	17
kurs_spot	spot exchange rate	160	18
kurs_tengah_bi	Bank Indonesia middle rate	74	9
laba	profit	57	7
laba_berjalan	current year profit	24	3
laba_bersih	net profit	216	24
laba_bersih_per_saham	earnings per share	34	4
laba_ditahan	retained earnings	15	2
laba_kotor	gross profit	118	14
laba_operasional	operational earnings	15	2
laba_pemilik_entitas_induk	attributable profit to parent entity	42	5
laba_penjualan_aset	asset sales revenue	9	2
ldr	loan-to-deposit ratio	14	2

lelang	auctions	16	2
liabilitas	liability	170	20
lq45	stock market index for the Indonesia Stock Exchange	98	12
luas	area	22	3
margin_ebitda	margin of earnings before interest, taxes, depreciation and amortization	8	2
market_share	market share	27	4
modal	capital	58	7
modal_dasar_inti_disetor	basic capital	101	12
modal_ekuitas	equity	96	11
modal_kerja	working capital	15	2
msci	The MSCI China index	26	4
nasdaq	Nasdaq stock market	559	63
net_profit_margin	net profit margin	18	3
nikkei	Nikkei Stock Average	302	34
nilai_akuisisi	acquisition value	18	3
nilai_ekspor	export value	98	11
nilai_impor	import value	10	2
nilai_investasi	investment value	55	7
nilai_kartu_kredit	credit card value	10	2
nilai_kontrak	contract value	39	5
nilai_kpr_kpa	mortgage value	20	3
nilai_kredit	credit value	32	4
nilai_kredit_umkm	SME credit value	12	2
nilai_transaksi	transaction value	192	22
nim	net interest margin	17	3
npf	non performing financing	17	3
obligasi	bond	93	11
pajak_bea_cukai	tax and customs	57	7
pasokan	supply	35	5
payout_ratio	payout ratio	77	9
pbt	profit before tax	26	3
pbv	Price by Volume Chart	26	3
pdb	Gross domestic product	55	7
pe_ratio	price earning ratio	63	8
pembayaran	payment	44	5
pembelian	purchase	9	1
pembelian_bersih	net buy	763	86
pembiayaan	financing	104	12
pemesanan	order	14	2
penawaran	offering	56	7
pendapatan	revenue	236	27
pendapatan_bersih	net revenue	121	14

pendapatan_negara	national revenue	17	3
pendapatan_pra_penjualan	pre-sales revenue	19	3
penempatan_dana	fund allocation	20	3
penempatan_dana_deposito	fund allocation in deposito	13	2
pengangguran	unemployment	20	3
penjualan	sales	171	20
penjualan_bersih	net sales	577	66
penyaluran_kredit	credit distribution	42	5
pertumbuhan_ekonomi	economic growth	27	4
piutang	accounts receivable	15	2
pmi	purchasing manager's index	25	3
populasi	population	17	3
portofolio	portfolio	35	4
portofolio_kredit	credit portfolio	19	3
premi	premium	13	2
premi_asuransi	insurance premium	33	4
probabilitas	probability	13	2
produksi	production	137	16
proyeksi_inflasi	inflation outlook	62	7
proyeksi_pdb	GDP outlook	26	3
proyeksi_pertumbuhan_ekonomi	economic growth outlook	54	7
rasio_casa	current account savings account ratio	10	2
rasio_modal	capital ratio	19	3
rasio_utang	loan ratio	11	2
rbc	risk based capital	17	3
resistance	resitance	114	13
restrukturisasi_kredit_pembiayaan	credit and financing restructurisation	13	2
return	return	17	2
return_reksadana	mutual funds returns	12	2
roa	return on assets	9	2
roe	return on equity	15	2
rugi	loss	170	19
saham	stocks	95	11
selisih	gap	12	2
selisih_kurs	exchange rate gap	17	2
shanghai	Shanghai composite index	247	28
sp_500	The Standard and Poor's 500	621	70
stoxx_600	The STOXX Europe 600 stock index	74	9
straits_times	Singapore Straits Times Index	260	29
suku_bunga	interest rate	69	8
suku_bunga_bank_indonesia	Bank Indonesia rate	30	4
suku_bunga_deposito	deposito interest rate	112	13
suku_bunga_kpr	mortgage interest rate	10	2

suku_bunga_tabungan	savings interest rate	18	3
suku_bunga_the_fed	The Fed interest rate	134	16
support	support	107	12
surat_utang	Letter of Credit	17	3
surplus_neraca_perdagangan	balance of trade surplus	11	2
surplus_transaksi_berjalan	current account surplus	8	2
tabungan	savings	26	4
taiex	Taiwan Stock Market Index	11	2
target_dana	funding target	34	4
target_kontrak	contract target	10	2
target_laba	profit target	16	2
target_pendapatan	revenue target	51	6
target_penjualan	sales target	18	3
target_produksi	production target	17	3
tpv	Total Portfolio Value	15	2
transaksi_bursa	Stock Market Transaction	24	3
transaksi_harian	Daily transaction	40	5
transaksi_saham	Stock Exchange	144	17
uang_muka	Down Payment	8	2
uang_pertanggungan	insurance	10	2
utang	Debt	30	4
valuasi	valuation	25	3
valuasi_saham	stock valuation	10	1
volume_penjualan	sales volume	18	3
volume_penjualan_batubara	coal sales volume	29	4
volume_penjualan_cpo	cpo sales volume	15	2
volume_penjualan_komoditas_lain	other commodities sales volume	18	3
volume_penjualan_semen	cement sales volume	11	2
volume_transaksi	transaction volume	30	5
volume_transaksi_bursa	stock market transaction volume	22	3
volume_transaksi_saham	stocks transaction volume	26	4
waran	warrant	15	2
yield	yield	24	3
yield_dividen	dividend yield	61	7
yield_obligasi	bond yield	130	15
yield_sbn	SBN yield	231	26
yield_sun	SUN yield	26	4
yield_us_treasury	US Treasury yield	215	25

B. THE CONFIGURATION OF MODELS AND SYSTEMS

SUBMODELS

Submodels		Targets
mtp_model1	BERT + convolutional	aspect1
mtp_model2	BERT + convolutional	aspect2
mtp_model2_0	BERT + convolutional	having_aspect2
mtp_model2_1	BERT + convolutional	aspect2_without_none
mtp_model3	BERT + convolutional	change
softmax_top3	BERT + convolutional + top-3 softmax	(aspect1, aspect2, change)
indonesianT5_model1	Pre-trained T5	aspect1
indonesianT5_model2	Pre-trained T5	aspect2
indonesianT5_model3	Pre-trained T5	change

EXTRACTION SYSTEMS

Systems	Submodels	Targets
Baseline1	IndonesianT5_model1	aspect1
	IndonesianT5_model2	aspect2
	IndonesianT5_model3	change
Baseline2	softmax_top3	(aspect1, aspect2, change)
NEAE-1	mtp_model1	aspect1
	mtp_model2	aspect2
	mtp_model3	change
NEAE-1	mtp_model1	Aspect1
	mtp_model2_0	having_aspect2
	mtp_model2_1	aspect2_without_none
	mtp_model3	change

HYPERPARAMETERS

Hyperparameters	Value
Optimizer	Adam
Learning rate	2e-05
Batch size	10
Maximum input length	512