# CHAPTER 4
# ANALYSIS AND DESIGN

## 4.1 Analysis

This study will use data from Company X. The data used is the answer data from several recruitment interview test questions that represent each aspect. Based on these answers, they can determine whether people for example from certain aspects are included in the category KS, K, C, B, or BS.

To get the answer, this study is going through 1 question for each personality aspect and from each answer there are criteria that meet whether the answers fall into the category KS, K, C, B, or BS.

Table 4.1: Interview Questions Table

| No | Aspect | Definition | Question |
|---|---|---|---|
| 1 | Motivation | Self encouragement to work / choose job position. | Kenapa Anda memilih bekerja di perusahaan ini? |
| 2 | Work Enthusiasm | Readiness in facing challenges at work. | Apakah anda siap bekerja shift, lembur dan dipindah-pindah bahkan ke luar pulau? |
| 3 | Self-awareness | The ability to introspect yourself and receive advice and input from others. | Apabila Anda dimarahi atasan karena dianggap salah. Apa yang akan anda lakukan? |

After getting the answer from the each questions, the next step is analyzing the answers from each questions that represent each aspect so people can tell which category does the answer belong to.

*Note : For this study, KS can be viewed as 1, K can be viewed as 2, C 3, B 4, and BS 5.*

Table 4.2: Answer Criteria for Motivation's Interview Question

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **KS** | **K** | **C** | **B** | **BS** |
| Originally register to a company because that company opens vacancies. | Register because he/she is joining someone else. | Join others but also already find out about the company. | Already find out what companies want to apply and indeed choose because he/she wants a career there. | Having prepared themselves beforehand even to find out the match between the values and characteristics of the work to be applied for. |

Table 4.3: Answer Criteria for Work Enthusiasm's Interview Question

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **KS** | **K** | **C** | **B** | **BS** |
| Not ready. | Depending on the situation / if only necessary. | Ready to shift and overtime but not moved outside the domicile. | Ready as long as you get the appropriate benefits. | Ready for all company needs. |

Table 4.4: Answer Criteria for Self-awareness's Interview Question

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **KS** | **K** | **C** | **B** | **BS** |
| Resign / quit / no interest in working. | Trying to survive as long as he feels he is right. | Just listen, accepted it because of he/she is your boss. | Apologize and promise not to do it again. | Apologize and take responsibility and try to resolve / correct the mistakes made. |

Here are some answers examples:

Table 4.5: Answer Examples from Motivation's Interview Question

| Answer | Category |
|---|---|
| Kalau saya lihat kemarin di depan ada pasang lowongan Pak jadi saya coba-coba saja. | KS |
| Saya direkomendasikan sama kakak saya soalnya kakak saya bekerja di perusahaan ini. | K |
| Saya mendaftar gara-gara saudara saya yang memberi rekomendasi kepada saya Pak untuk masuk perusahaan ini, tetapi saya juga sudah cari-cari info sih mengenai perusahaan. | C |
| Saya sudah lama kagum dengan perusahaan Pak karena perkembangannya cepat. Saya senang karena lowongannya ada yang saya kuasai jadi saya langsung daftar. | B |
| Saya lulusan SMK TKJ Pak saya sudah ada rencana kerja di network, kemarin waktu lihat lowongan di sini saya rasa cocok dengan kemampuan saya sehingga saya tertarik Pak. | BS |

Table 4.6: Answer Examples from Work Enthusiasm's Interview Question

| Answer | Category |
|---|---|

| | |
|---|---|
| Saya belum bisa kasih jawaban sekarang karena banyak yang perlu saya pertimbangkan Pak. | KS |
| Tergantung situasi sih Pak, jika perlu maka saya ok ok saja. | K |
| Saya siap Pak, tetapi kalau bisa saya jangan ditempatkan di luar pulau ya Pak. | C |
| Saya bersedia jika ada gaji lembur yang sepadan. | B |
| Siap semuanya Pak. | BS |

Table 4.7: Answer Examples from Self-awareness's Interview Question

| **Answer** | **Category** |
|---|---|
| Jika hal ini bukan karena kesalahan saya dan berulang-ulang secara terus menerus maka saya akan mengundurkan diri. | KS |
| Jika saya merasa diri saya tidak bersalah namun dianggap salah ya saya berusaha untuk meluruskan. | K |
| Ya saya dengarkan saja, mungkin dari itu bisa jadi masukan untuk evaluasi diri saya Pak. | C |
| Saya langsung minta maaf dan berjanji tidak mengulanginya lagi. | B |
| Saya bertanggung jawab dan meminta maaf kepada atasan saya. | BS |

So from the example above, for example when asked Motivation's Interview Question *"Kenapa Anda memilih bekerja di perusahaan ini?"*. If the person answers *"Kalau saya lihat kemarin di depan ada pasang lowongan Pak jadi saya coba-coba saja."*, it will go into the KS category in motivation. Then if the person answers *"Saya direkomendasikan sama kakak saya soalnya kakak saya*

*bekerja di perusahaan ini."*, it will go to K category in motivation and etc. To determine this answer will go in which category is determined by the criteria that already explained in table 4.2, 4.3, and 4.4.

Now to create a classification program that can classify some of the answers to get the value of each aspect (KS, K, C, B, and BS), this study will use the LDA algorithm. But what is used is a supervised version that provides a label before the data is clustered or grouped [4].

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of several documents commonly referred to as a corpus [16]. The document here usually consists of text. LDA is a topic modeling method which means a method that can be used to find several abstract topics in collection of documents [17]. The explanation about LDA below will be based on Blei et al's study [13] and Thushan Ganegedara's explanation [3]. But, most importantly it is based on Daniel Ramage et al's study about L-LDA [4] and Yiqi Bai and Jie Wang's study [14].

Before going to the L-LDA like stated before, the LDA pictures a document is created by the distribution of topics (θ) and a topic created by the distribution of words (β) determined by two dirichlet distribution parameters α and η.

In each document, the LDA views each document as a collection of words or bag of words $\boldsymbol{w}^{(d)}$.

$\boldsymbol{w}^{(d)} = (w_1, \ldots, w_{Nd})$

$d$ = document $d$ , $w$ = word, $N_d$ = total of words in document $d$

The collection of documents is called a corpus ($D$) containing $D = \{ \boldsymbol{w}^{(1)}, \ldots , \boldsymbol{w}^{(M)} \}$ where $M$ is the total number of documents and each word is added to a vocabulary collection ($V$) . where $V = \{ w_1, w_2, \ldots, w_T \}$.

So we have the following :

- α: The dirichlet distribution variable that controls the topic distribution.
- η: The dirichlet distribution variable that control the word distribution.
- K: Number of topics.
- k: A topic.
- M: Number of documents.
- d: A document.
- D: Document corpus.
- V: Vocabulary or vocabulary size.
- T: Number of words or terms in the corpus
- $N_d$: Total words in each document d.
- w: Word in the document.
- **w**: Describe a document d with a total of $N_d$ words.
- z: Topic from a topic dictionary.
- θ: Distribution of topics in the document.
- β: Distribution of words in the topic.

From the definition above, to get distribution of topics (θ), a matrix M × K will be created and the contents will be the total number of occurrences of each topic in each document (Document Topic Matrix). Because later, each word in the document can refer to certain topics, then in the document will look for the total appearance of topics in the words in the document. Then to get the distribution of words *(*β), a matrix of size K × T will be created (Topic Term Matrix). The contents of the matrix are the total occurrences of words in a particular topic. So the corpus will count the total words that refer to the topic.

Therefore in this study, the topic is a category in each aspect and the document is the answer to the text of each question in a particular aspect. So K = 5, because there are 5 categories, already explained earlier per each aspect. (KS, K, C, B, BS) or (1, 2, 3, 4, 5).

To understand the flow of this algorithm, we will explain how this algorithm works based on implementation of L-LDA algorithm by Jiahong Zhou [18].

To continue, it is necessary to create a word dictionary and also a topic. And later each word in the document will be replaced by the word number or word id. And a topic in a document will be replaced by topic number or topic id.

Table 4.8: Topic Vocabulary

| Topic id | Topic | |
|---|---|---|
| 0 | KS | 1 |
| 1 | K | 2 |
| 2 | C | 3 |
| 3 | B | 4 |
| 4 | BS | 5 |

Because the total number of documents and words used is too much for, so it is not possible to put the dictionary here. So examples will be given for only one document. Example like “Siap Semua Pak”. Then the word or term vocabulary will look like :

Table 4.9: Word or Term Vocabulary Example

| Word or term id | Word or term |
|---|---|
| 0 | siap |
| 1 | semua |
| 2 | pak |

After that in the document, the words or terms inside the document will be replaced with the word id or term id so that it becomes *"0, 1, 2"*. Another example is when a document has text *“Semua Siap Pak”*, it will be changed into *“1, 0, 2”*.

We are going to do this step throughout the training data. So we've got:

- M = 250,
- K = 5,
- T = (depends on the aspect),

- $T_{motivation}$ = 371,
- $T_{work\ enthusiasm}$= 206,
- $T_{self\text{-}awareness}$ = 230

Then we can create matrix that will help to count θ and β for each aspect based on the explanation above. *(The matrix will be so huge, so we will not include the matrix content here)*. But before creating those matrix, there is one thing that is missed. Because we are using L-LDA algorithm. We will create 1 more matrix indicating which documents will go into which cluster or group. So we need to give every document a label or maybe more (Λ). The matrix is called Lambda or we can symbolize it as $L_{(d)}$. The size of this matrix is M × K and the content inside is a indicator which document belongs to topics.

So for example, if we have a corpus consists of :

Table 4.10: Corpus Examples

| Doc no | Answer | Category |
|---|---|---|
| 1 | Kalau saya lihat kemarin di depan ada pasang lowongan Pak jadi saya coba-coba saja. | KS |
| 2 | Saya direkomendasikan sama kakak saya soalnya kakak saya bekerja di perusahaan ini. | K |
| 3 | Saya mendaftar gara-gara saudara saya yang memberi rekomendasi kepada saya Pak untuk masuk perusahaan ini, tetapi saya juga sudah cari-cari info sih mengenai perusahaan. | C |
| 4 | Saya sudah lama kagum dengan perusahaan Pak karena perkembangannya cepat. Saya senang karena lowongannya ada yang saya kuasai jadi saya langsung daftar. | B |
| 5 | Saya lulusan SMK TKJ Pak saya sudah ada rencana kerja di network, kemarin waktu lihat lowongan di sini saya rasa cocok dengan kemampuan saya sehingga saya tertarik Pak. | BS |

The Lambda matrix will look like :

Table 4.11: Lambda Matrix Based on Table 4.10

| | **0**<br>**KS** | **1**<br>**K** | **2**<br>**C** | **3**<br>**B** | **4**<br>**BS** |
|---|---|---|---|---|---|
| **Doc 1** | 1 | 0 | 0 | 0 | 0 |
| **Doc 2** | 0 | 1 | 0 | 0 | 0 |
| **Doc 3** | 0 | 0 | 1 | 0 | 0 |
| **Doc 4** | 0 | 0 | 0 | 1 | 0 |
| **Doc 5** | 0 | 0 | 0 | 0 | 1 |

Each matrix index represents each ids, So when we look at table 4.8, we see that KS has an id, it is 0, K is 1, C is 2, B is 3, and BS is 4. So in the matrix above, column index represents each topic id. So column index 0 is KS, 1 is K, 2 is C, 3 is B, and 4 is BS. And row index represents document number. Index 0 is document 1, 1 is document 2, 2 is document 3, 3 is document 4, and 4 is document 5. Now we only need to mark if document 1 goes to the KS category or to the KS cluster then we mark 1, if not will be marked with 0. So if 1 document can go to 2 clusters, we just need to mark if that document will be in 2 cluster, Example :

Table 4.12: Lambda Matrix Example

| | **A** | **B** | **C** | **D** | **E** |
|---|---|---|---|---|---|
| **Doc 1** | 1 | 1 | 0 | 0 | 0 |

Like stated in chapter 1 and 2, in this study we want to compare how if we group the data that first contains the common word or neutral word and the second is directly clustered or grouped or does not care about the common word or neutral word in the document, so if you use the second approach, Lambda will be like in the table 4.11. If you use the neutral or common word then we will add one topic or one category or one more cluster which is called neutral or common to select some words that do not refer to topics or categories KS, K, C, B, and BS. But go to the cluster or category neutral or common. Therefore from that, the

topic of vocabulary will be added by one cluster named neutral or common. Then we will mark that every document has a neutral or common words. So a document will consist of words that represent a topic and do not represent any topic (common).

Table 4.13: Topic Vocabulary with Common or Neutral Topic

| Topic id | Topic | |
|---|---|---|
| 0 | common | |
| 1 | KS | 1 |
| 2 | K | 2 |
| 3 | C | 3 |
| 4 | B | 4 |
| 5 | BS | 5 |

Table 4.14: Lambda Matrix Based on Table 4.10 with Common Topic.

| | **0** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|---|
| | **common** | **KS** | **K** | **C** | **B** | **BS** |
| **Doc 1** | 1 | 1 | 0 | 0 | 0 | 0 |
| **Doc 2** | 1 | 0 | 1 | 0 | 0 | 0 |
| **Doc 3** | 1 | 0 | 0 | 1 | 0 | 0 |
| **Doc 4** | 1 | 0 | 0 | 0 | 1 | 0 |
| **Doc 5** | 1 | 0 | 0 | 0 | 0 | 1 |

Now, we will proceed to the next step, we will create Document Topic Matrix and Topic Term Matrix for each aspect. To create those matrices. The initial step to do is to stick to all the words in each document a random topic in accordance with Lambda earlier. So for example document 1 is given directions in Lambda that it falls into the common and KS categories. Then later on every word in document 1 will be given a random topic that is common or KS. We will do it for the entire document and for each word or terms in a document.

Because the matrix is very large, to simplify and clarify how it works, we will use a simple example. For example we have 3 documents :

Table 4.15: Example Corpus

| No | Dokumen | Label |
|---|---|---|
| 1 | I like cat | cat |
| 2 | I like dog | dog |
| 3 | I like bird | bird |

We will create a vocabulary dictionary and topic vocabulary dictionary based on those 3 documents. And this example will use the first approach, a document has a common or neutral word.

Table 4.16: Topic Vocabulary Based on Table 4.15

| id | topic |
|---|---|
| 0 | common |
| 1 | cat |
| 2 | dog |
| 3 | bird |

Table 4.17: Word or Term Vocabulary Based on Table 4.15

| Word or term id | Word or term |
|---|---|
| 0 | i |
| 1 | like |
| 2 | cat |
| 3 | dog |
| 4 | bird |

Table 4.18: Lambda Based on Table 4.15

| | 0<br>common | 1<br>cat | 2<br>dog | 3<br>bird |
|---|---|---|---|---|
| **Doc 1** | 1 | 1 | 0 | 0 |
| **Doc 2** | 1 | 0 | 1 | 0 |
| **Doc 3** | 1 | 0 | 0 | 1 |

Then we can replace the words inside all the documents with their ids based on the vocabulary dictionary and then we will add a topic in each word based on their Lambda.

Table 4.19: Replace Each Words Inside Documents With Their Ids and Add a Topic in Every Word

| **Doc No** | | | |
|---|---|---|---|
| Doc 1 | i | like | cat |
| Replaced By | 0 | 1 | 2 |
| Label or Topic | 0 | 1 | 0 |
| | | | |
| Doc 2 | i | like | dog |
| Replaced By | 0 | 1 | 3 |
| Label or Topic | 2 | 2 | 2 |
| | | | |
| Doc 3 | i | like | bird |
| Replaced By | 0 | 1 | 4 |
| Label or Topic | 3 | 3 | 0 |

From table 4.19, we will create Document Topic Matrix for counting θ and Topic Term Matrix for counting β. The matrix content is same like explained before.

Table 4.20: Document Topic Matrix Based on Table 4.19

| | | **0** | **1** | **2** | **3** |
|---|---|---|---|---|---|
| | | **common** | **cat** | **dog** | **bird** |
| **0** | **Doc 1** | 2 | 1 | 0 | 0 |
| **1** | **Doc 2** | 0 | 0 | 3 | 0 |
| **2** | **Doc 3** | 1 | 0 | 0 | 2 |

Row index represents document id, and column index represents topic id.

Table 4.21: Topic Term Matrix Based on Table 4.19

| | | **0** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|---|
| | | **i** | **like** | **cat** | **dog** | **bird** |
| **0** | **common** | 1 | 0 | 1 | 0 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **1** | **cat** | 0 | 1 | 0 | 0 | 0 |
| **2** | **dog** | 1 | 1 | 0 | 1 | 0 |
| **3** | **bird** | 1 | 1 | 0 | 0 | 0 |

Row index represents topic id and column index represents word or term id.

After getting those two matrices, getting θ and β will not be possible because the matrix data is not normalized (not accurate). It will result an inaccurate result. So, now we are going to train the model by slowly change each topic in each word inside documents by using collapsed Gibbs sampling. Like Daniel Ramage et al stated in their research [4], but originally comes from Griffiths and Steyvers in 2004 [19] :

$$P(z_i = k | z_{-i}) = \frac{n_t^{(k,-i)} + \eta^{(k,t)}}{\sum_{t=1}^{T} n_t^{(k,-i)} + \eta^{(k,t)}} \times \frac{n_k^{(d,-i)} + \alpha^{(d,k)}}{\sum_{k=1}^{K} n_k^{(d,-i)} + \alpha^{(d,k)}}$$

Where :

- $n^{(k)}_t$ = Total count of word or term t in topic k
- $\eta^{(k,t)}$ = Distribution variable of word or term t in topic k
- $n^{(d)}_k$ = Total count of topic k in document d
- $\alpha^{(d,k)}$ = Distribution variable of topic k in document d

The -i indicates that the total count are not the present condition. ($n^{(k)}_t$ - 1 and $n^{(d)}_k$ - 1)

How do we get the α & η values? Based on Yiqi Bai and Jie Wang's research, they stated that it is conventional to set the values [14]:

- α = 50 / K
- η = 0.1

So, from table 4.20 and 4.21. We will slowly change the topic in each word inside documents. Starting from document 1 that has :

Table 4.22: Current Condition of Document 1

| **Doc No** | | | |
|---|---|---|---|
| Doc 1 | i | like | cat |
| Doc id | 0 | 1 | 2 |
| Label or Topic | 0 | 1 | 0 |

We will change the topic in word I in document 1. So we will decrement the count of word I in document 1 and common topic inside Document Topic Matrix then in common topic and term i inside Topic Term Matrix.

Table 4.23: Current Condition of Document Topic Matrix

| | | **0** | **1** | **2** | **3** |
|---|---|---|---|---|---|
| | | **common** | **cat** | **dog** | **bird** |
| **0** | **Doc 1** | 1 | 1 | 0 | 0 |
| **1** | **Doc 2** | 0 | 0 | 3 | 0 |
| **2** | **Doc 3** | 1 | 0 | 0 | 2 |

Table 4.24: Current Condition of Topic Term Matrix

| | | **0** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|---|
| | | **i** | **like** | **cat** | **dog** | **bird** |
| **0** | **common** | 0 | 0 | 1 | 0 | 1 |
| **1** | **cat** | 0 | 1 | 0 | 0 | 0 |
| **2** | **dog** | 1 | 1 | 0 | 1 | 0 |
| **3** | **bird** | 1 | 1 | 0 | 0 | 0 |

Then we will apply the formula above (In this example, we will use $\eta$ = 0.001 and $\alpha$ = 0.01). But remember that we have a Lambda matrix that indicates the topic occurrence inside the documents. So the distribution of topics should be :

$$\alpha^{(d)} = L^{(d)} \times \alpha$$

Table 4.25: $\alpha^{(d)}$ Based on the Example

| **0** | **1** | **2** | **3** |
|---|---|---|---|
| **common** | **cat** | **dog** | **bird** |

| | | | | |
|---|---|---|---|---|
| **Doc 1** | 0,01 | 0,01 | 0 | 0 |
| **Doc 2** | 0,01 | 0 | 0,01 | 0 |
| **Doc 3** | 0,01 | 0 | 0 | 0,01 |

Table 4.26: Formula Calculation (Left Side)

| | |
|---|---|
| $\text{Left}^{(\text{topic common word i})} =$ | $\frac{0+0{,}001}{(0{,}001+0{,}001+1{,}001+0{,}001+1{,}001)} = \frac{0{,}001}{2{,}005} = 0{,}000498753$ |
| $\text{Left}^{(\text{topic cat, word i})} =$ | $\frac{0+0{,}001}{(0{,}001+1{,}001+0{,}001+0{,}001+0{,}001)} = \frac{0{,}001}{1{,}005} = 0{,}000995025$ |
| $\text{Left}^{(\text{topic dog, word i})} =$ | $\frac{1+0{,}001}{(1{,}001+1{,}001+0{,}001+1{,}001+0{,}001)} = \frac{1{,}001}{3{,}005} = 0{,}333111481$ |
| $\text{Left}^{(\text{topic bird, word i})} =$ | $\frac{1+0{,}001}{(1{,}001+1{,}001+0{,}001+0{,}001+0{,}001)} = \frac{1{,}001}{2{,}005} = 0{,}49925187$ |

Table 4.27: Formula Calculation (Right Side)

| | |
|---|---|
| $\text{Right}^{(\text{doc 1, topic common})} =$ | $\frac{1+0{,}01}{1{,}01+1{,}01+0+0} = \frac{1{,}01}{2{,}02} = 0{,}5$ |
| $\text{Right}^{(\text{doc 1, topic cat})} =$ | $\frac{1+0{,}01}{1{,}01+1{,}01+0+0} = \frac{1{,}01}{2{,}02} = 0{,}5$ |
| $\text{Right}^{(\text{doc 1, topic dog})} =$ | $\frac{0+0}{1{,}01+1{,}01+0+0} = \frac{0}{2{,}02} = 0$ |
| $\text{Right}^{(\text{doc 1, topic bird})} =$ | $\frac{0+0}{1{,}01+1{,}01+0+0} = \frac{0}{2{,}02} = 0$ |

New Topic Probability for Word I :

Table 4.28: New Topic Probability Calculation For Word I

| Topic | (Left * Right) | Probability |
|---|---|---|
| common | 0,000498753 × 0.5 = 0,000249501 | 0,000249501/( 0,000249501+ 0,000249501 + 0 + 0) = **0,333998204** |
| cat | 0,000995025 × 0.5 = 0,000497512 | 0,000497512 /( 0,000249501+ 0,000249501 + 0 + 0) = **0,666001796** |
| dog | 0,333111481 × 0 = 0 | 0 /( 0,000249501+ 0,000249501 + 0 + 0) = **0** |
| bird | 0,49925187 × 0 = 0 | 0 /( 0,000249501+ 0,000249501 + 0 + 0) = **0** |

We will get a new topic (sample new topic) based on that probability. From it, we can see that there is a chance of **33,3998204 %** that the new topic selected will be a common topic for word I. But there is also another **66,6001796%** that the new topic selected will be a cat topic. Because the cat topic has a bigger chance, the new topic selected is a cat topic for the word I. After that we change the word I topic from the document.

Table 4.29: Current Condition of Document 1 (Changed I Topic to Cat)

| Doc No | | | |
|---|---|---|---|
| Doc 1 | i | like | cat |
| Doc id | 0 | 1 | 2 |
| Label or Topic | 1 | 1 | 0 |

Then update the Document Topic Matrix and Topic Term Matrix.

Table 4.30: Updated Condition of Document Topic Matrix Based on Table 4.29

| | | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| | | common | cat | dog | bird |
| 0 | Doc 1 | 1 | 2 | 0 | 0 |
| 1 | Doc 2 | 0 | 0 | 3 | 0 |
| 2 | Doc 3 | 1 | 0 | 0 | 2 |

Table 4.31: Updated Condition of Topic Term Matrix Based on Table 4.29

| | | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| | | i | like | cat | dog | bird |
| 0 | common | 0 | 0 | 1 | 0 | 1 |
| 1 | cat | 1 | 1 | 0 | 0 | 0 |
| 2 | dog | 1 | 1 | 0 | 1 | 0 |
| 3 | bird | 1 | 1 | 0 | 0 | 0 |

Do the same steps as the example for all words in each document. And If until the end of the document, repeat this step from document 1. The steps taken by this example will be the same for 250 data that has been prepared. Also do all these steps in all aspects. So we will have 3 classification models for each aspect.

After the model is created, then we can search for the distribution of topics in each document (θ) in the corpus and also search for word distribution in topics (β). To get θ & β we will use the formula provided by Griffiths and Steyvers [19].

To find θ :

$$\theta^{(d,k)} = \frac{n_k^{(d)} + \alpha^{(d,k)}}{\sum_{k=1}^{K} n_k^{(d)} + \alpha^{(d,k)}}$$

Where :

- $n^{(d)}_k$ = Total count of topic k in document d
- $\alpha^{(d, k)}$ = Distribution variable of topic k in document d

To find β :

$$\beta^{(k,t)} = \frac{n_t^{(k)} + \eta^{(k,t)}}{\sum_{t=1}^{T} n_t^{(k)} + \eta^{(k,t)}}$$

Where :

- $n^{(k)}_t$ = Total count of word or term t in topic k
- $\eta^{(k, t)}$ = Distribution variable of word or term t in topic k

For example, based from the example in table 4.15, the final model (the final Document Topic matrix and Topic Term Matrix) :

Table 4.32: Final Document Topic Matrix Based on Table 4.15

| | | **0** | **1** | **2** | **3** |
|---|---|---|---|---|---|
| | | **common** | **cat** | **dog** | **bird** |
| **0** | **Doc 1** | 2 | 1 | 0 | 0 |
| **1** | **Doc 2** | 2 | 0 | 1 | 0 |
| **2** | **Doc 3** | 2 | 0 | 0 | 1 |

Table 4.33: Final Topic Term Matrix Based on Table 4.15

| **0** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| **i** | **like** | **cat** | **dog** | **bird** |

| | | | | | | |
|---|---|---|---|---|---|---|
| **0** | **common** | 3 | 3 | 0 | 0 | 0 |
| **1** | **cat** | 0 | 0 | 1 | 0 | 0 |
| **2** | **dog** | 0 | 0 | 0 | 1 | 0 |
| **3** | **bird** | 0 | 0 | 0 | 0 | 1 |

Based on the formula above, θ =

Table 4.34: Distribution of Topics Inside Documents Based on Table 4.15

| Document | common | cat | dog | bird |
|---|---|---|---|---|
| Doc 1 | 2 + 0,01 / (2,01 + 1,01) = **0,6655629139072848** | 1 + 0,01 / (2,01 + 1,01) = **0,3344370860927152** | 0 + 0 / (2,01 + 1,01) = **0** | 0 + 0 / (2,01 + 1,01) = **0** |
| Doc 2 | 2 + 0,01 / (2,01 + 1,01) = **0,6655629139072848** | 0 + 0 / (2,01 + 1,01) = **0** | 1 + 0,01 / (2,01 + 1,01) = **0,3344370860927152** | 0 + 0 / (2,01 + 1,01) = **0** |
| Doc 3 | 2 + 0,01 / (2,01 + 1,01) = **0,6655629139072848** | 0 + 0 / (2,01 + 1,01) = **0** | 0 + 0 / (2,01 + 1,01) = **0** | 1 + 0,01 / (2,01 + 1,01) = **0,3344370860927152** |

So, Document 1 consists of **66,55629139072848 %** common topic (words that do not refer any topic or words that are general in nature which means the words also can sometimes appear frequently in each document.) and **33,44370860927152 %** cat topic, document 2 consists of **66,55629139072848 %** common topic and **33,44370860927152 %** dog topic, and etc.

And β =

Table 4.35: Distribution of Words or Terms Inside Topics Based on Table 4.15

| Topic | i | like | cat | dog | bird |
|---|---|---|---|---|---|
| common | 3 + 0,001/ (3,001 + 3,001 + 0,001 + 0,001 + 0,001) = **0,4997502081598668** | 3 + 0,001/ (3,001 + 3,001 + 0,001 + 0,001 + 0,001) = **0,4997502081598668** | 0 + 0,001/ (3,001 + 3,001 + 0,001 + 0,001 + 0,001) = **1,665278934221482e-4** | 0 + 0,001/ (3,001 + 3,001 + 0,001 + 0,001 + 0,001) = **1,665278934221482e-4** | 0 + 0,001/ (3,001 + 3,001 + 0,001 + 0,001 + 0,001) = **1,665278934221482e-4** |

| | | | | | |
|---|---|---|---|---|---|
| cat | 0 + 0,001 / (0,001 + 0,001 + 1,001 + 0,001 + 0,001) = **9,950248756218905e-4** | 0 + 0,001 / (0,001 + 0,001 + 1,001 + 0,001 + 0,001) = **9,950248756218905e-4** | 1 + 0,001 / (0,001 + 0,001 + 1,001 + 0,001 + 0,001) = **0,9960199004975124** | 0 + 0,001 / (0,001 + 0,001 + 1,001 + 0,001 + 0,001) = **9,950248756218905e-4** | 0 + 0,001 / (0,001 + 0,001 + 1,001 + 0,001 + 0,001) = **9,950248756218905e-4** |
| dog | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 1,001 + 0,001) = **9,950248756218905e-4** | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 1,001 + 0,001) = **9,950248756218905e-4** | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 1,001 + 0,001) = **9,950248756218905e-4** | 1 + 0,001 / (0,001 + 0,001 + 0,001 + 1,001 + 0,001) = **0,9960199004975124** | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 1,001 + 0,001) = **9,950248756218905e-4** |
| bird | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 0,001 + 1,001) = **9,950248756218905e-4** | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 0,001 + 1,001) = **9,950248756218905e-4** | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 0,001 + 1,001) = **9,950248756218905e-4** | 0 + 0,001 / (0,001 + 0,001 + 0,001 + 0,001 + 1,001) = **9,950248756218905e-4** | 1 + 0,001 / (0,001 + 0,001 + 0,001 + 0,001 + 1,001) = **0,9960199004975124** |

So, topic cat has word distributions : **9,950248756218905e-4 * i**, **9,950248756218905e-4 * like**, **0,9960199004975124 * cat**, **9,950248756218905e-4 * dog**, and **9,950248756218905e-4 * bird,** topic dog has word distributions : **9,950248756218905e-4 * i**, **9,950248756218905e-4 * like**, **9,950248756218905e-4 * cat**, **0,9960199004975124 * dog,** and **9,950248756218905e-4 * bird**, and etc.

We will make a model with 2 approaches that were explained earlier. The initial approach that uses the word common (common topic) and the approach that does not use the common word, directly grouped into certain categories. So in total there will be 6 models made.

The word distribution in this research case, from each aspect will look like this:

***Approach 1 : With common words***

*Remember that KS can be viewed as 1, K can be viewed as 2, C 3, B 4, and BS 5.*

Table 4.36: Words Distributions in Motivation

| topic | dibilangin | temen | coba | daftar | sini | kata | cocok | banget | kemarin | tarik | bapak | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| common | 8.318775476249E-05 | 8.318775476249E-05 | 0.000915065 3 | 0.060810248 7 | 0.056650861 | 0.008401963 2 | 0.015888612 | 0.003410697 9 | 0.013393228 5 | 0.026703269 3 | 0.005906330 6 | ... |
| 1 | 0.0005022602 | 0.0005022602 | 0.110999497 7 | 0.0005022602 | 0.0005022602 | 0.0005022602 | 0.0005022602 | 0.0005022602 | 0.010547463 6 | 0.0005022602 | 0.0005022602 | ... |
| 2 | 0.0071162318 | 0.0037275 5 | 0.024059640 8 | 0.0037 2755 | 0.047780413 4 | 0.030837004 4 | 0.0003388682 | 0.0037 2755 | 0.020670959 | 0.0003388682 | 0.020670959 | ... |
| 3 | 0.0005073567 | 0.0055809234 | 0.025875190 3 | 0.0005073567 | 0.0005073567 | 0.0005073567 | 0.0005073567 | 0.0005073567 | 0.0005073567 | 0.025875190 3 | 0.0005073567 | ... |
| 4 | 0.0002314279 | 0.0002314279 | 0.0002314279 | 0.0002314279 | 0.0002314279 | 0.0002314279 | 0.002545707 | 0.0002314279 | 0.004859986 1 | 0.034945614 4 | 0.0002314279 | ... |
| 5 | 0.0006169031 | 0.0006169031 | 0.000616903 1 | 0.000616903 1 | 0.000616903 1 | 0.000616903 1 | 0.000616903 1 | 0.000616903 1 | 0.000616903 1 | 0.012954966 1 | 0.012954966 1 | ... |

Table 4.37: Words Distributions in Work Enthusiasm

| topic | lembur | hampir | tiap | hari | tidak | sedia | siap | resiko | orang | kerja | keluarga | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| common | 0.0720907928 | 0.0001598465 | 0.0001598465 | 0.0001598465 | 0.0976662404 | 0.0225383632 | 0.1887787724 | 0.0001598465 | 0.0065537084 | 0.0401214834 | 0.0225383632 | ... |
| 1 | 0.0008503401 | 0.0008503401 | 0.0008503401 | 0.0008503401 | 0.0008503401 | 0.0008503401 | 0.0008503401 | 0.0008503401 | 0.0348639456 | 0.0008503401 | 0.0008503401 | ... |
| 2 | 0.0006345178 | 0.0069796954 | 0.0323604061 | 0.0577411168 | 0.0069796954 | 0.0006345178 | 0.0006345178 | 0.0006345178 | 0.0006345178 | 0.0006345178 | 0.0006345178 | ... |
| 3 | 0.0004139073 | 0.0004139073 | 0.0004139073 | 0.0004139073 | 0.0004139073 | 0.0004139073 | 0.0004139073 | 0.0004139073 | 0.0418046358 | 0.0004139073 | 0.0004139073 | ... |
| 4 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | 0.0009 65251 | ... |
| 5 | 0.00051 12474 | 0.00051 12474 | 0.00051 12474 | 0.00051 12474 | 0.00051 12474 | 0.0260 736196 | 0.2816 973415 | 0.0158 486708 | 0.00051 12474 | 0.00051 12474 | 0.00051 12474 | ... |

Table 4.38: Words Distributions in Self-awareness

| topic | langsung | minta | maaf | atas | tegur | salah | tanggung | jawab | marah | alas | jelas | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| common | 0.0036427732 | 0.0341950646 | 0.0341950646 | 0.0858989424 | 0.0048178613 | 0.1493537015 | 0.0059929495 | 0.0095182139 | 0.0330199765 | 0.0071680376 | 0.0083431257 | ... |
| 1 | 0.0008064516 | 0.0008064516 | 0.0008064516 | 0.0088709677 | 0.0008064516 | 0.0008064516 | 0.0008064516 | 0.0008064516 | 0.1056451613 | 0.0330645161 | 0.0008064516 | ... |
| 2 | 0.0005102041 | 0.0362244898 | 0.0209183673 | 0.0005102041 | 0.0005102041 | 0.0362244898 | 0.0005102041 | 0.0005102041 | 0.0005102041 | 0.0005102041 | 0.1229591837 | ... |
| 3 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0004694836 | 0.0098591549 | ... |
| 4 | 0.0163 | 0.2350 | 0.2254 | 0.0003 | 0.0003 | 0.0067 | 0.0003 | 0.0003 | 0.0067 | 0.0003 | 0.0003 | ... |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 987138 | 482315 | 019293 | 215434 | 215434 | 524116 | 215434 | 215434 | 524116 | 215434 | 215434 | |
| 5 | 0.0091 666667 | 0.1341 666667 | 0.1508 333333 | 0.0008 333333 | 0.0008 333333 | 0.0008 333333 | 0.1591 666667 | 0.1508 333333 | 0.0008 333333 | 0.0008 333333 | 0.0008 333333 | ... |

## ***Approach 2 : Without common words***

Table 4.39: Words Distributions in Motivation

| topic | dibilangin | temen | coba | daftar | sini | kata | cocok | banget | kemarin | tarik | bapak | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00025 56891 | 0.00025 56891 | 0.05650 72871 | 0.03860 90514 | 0.02838 14881 | 0.00536 94707 | 0.00281 25799 | 0.00025 56891 | 0.01559 7034 | 0.00792 63615 | 0.00025 56891 | ... |
| 2 | 0.00373 599 | 0.00195 69472 | 0.01441 02473 | 0.04109 58904 | 0.05177 01477 | 0.02152 64188 | 0.00551 50329 | 0.00551 50329 | 0.01796 8333 | 0.01618 92902 | 0.01263 12044 | ... |
| 3 | 0.00025 31005 | 0.00278 41053 | 0.01290 81245 | 0.02809 41534 | 0.03062 51582 | 0.00784 61149 | 0.00278 41053 | 0.00278 41053 | 0.00784 61149 | 0.03062 51582 | 0.00278 41053 | ... |
| 4 | 0.00012 03225 | 0.00012 03225 | 0.00012 03225 | 0.02177 8366 | 0.03020 09385 | 0.00132 35471 | 0.01215 25689 | 0.00132 35471 | 0.00733 96703 | 0.03140 41632 | 0.00493 3221 | ... |
| 5 | 0.00036 88676 | 0.00036 88676 | 0.00036 88676 | 0.02618 95979 | 0.01881 22464 | 0.00405 75433 | 0.01881 22464 | 0.00036 88676 | 0.00405 75433 | 0.01512 35706 | 0.01143 48949 | ... |

Table 4.40: Words Distributions in Work Enthusiasm

| topic | lembur | hampir | tiap | hari | tidak | sedia | siap | resiko | orang | kerja | keluarga | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.03792 13483 | 0.00046 81648 | 0.00046 81648 | 0.00046 81648 | 0.06132 9588 | 0.00983 14607 | 0.03792 13483 | 0.00046 81648 | 0.02855 80524 | 0.02387 64045 | 0.00983 14607 | ... |
| 2 | 0.04539 15454 | 0.00381 15038 | 0.01767 15177 | 0.03153 15315 | 0.06964 65696 | 0.01420 65142 | 0.07657 65766 | 0.00034 65003 | 0.00034 65003 | 0.01074 15107 | 0.00727 65073 | ... |
| 3 | 0.02521 21817 | 0.00024 96256 | 0.00024 96256 | 0.00024 96256 | 0.05267 09935 | 0.00773 83924 | 0.08012 98053 | 0.00024 96256 | 0.02770 84373 | 0.01023 4648 | 0.01772 34149 | ... |
| 4 | 0.02460 98439 | 0.00060 02401 | 0.00060 02401 | 0.00060 02401 | 0.01260 5042 | 0.00660 26411 | 0.13865 54622 | 0.00060 02401 | 0.00060 02401 | 0.00060 02401 | 0.00660 26411 | ... |
| 5 | 0.02872 58248 | 0.00028 44141 | 0.00028 44141 | 0.00028 44141 | 0.01734 92605 | 0.02588 16837 | 0.25056 88282 | 0.00881 68373 | 0.00312 85552 | 0.03725 8248 | 0.00597 26962 | ... |

Table 4.41: Words Distributions in Self-awareness

| topic | langsung | minta | maaf | atas | tegur | salah | tanggung | jawab | marah | alas | jelas | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00044 44444 | 0.02711 11111 | 0.00933 33333 | 0.04933 33333 | 0.00488 88889 | 0.036 | 0.00044 44444 | 0.00044 44444 | 0.08488 88889 | 0.01822 22222 | 0.01822 22222 | ... |
| 2 | 0.00023 75297 | 0.03111 63895 | 0.02874 10926 | 0.04061 75772 | 0.00023 75297 | 0.11187 64846 | 0.00261 28266 | 0.00498 81235 | 0.01923 9905 | 0.00261 28266 | 0.05961 99525 | ... |
| 3 | 0.00286 45833 | 0.00546 875 | 0.00546 875 | 0.04973 95833 | 0.00286 45833 | 0.04713 54167 | 0.00286 45833 | 0.00286 45833 | 0.02369 79167 | 0.00807 29167 | 0.00807 29167 | ... |
| 4 | 0.01037 41497 | 0.14472 78912 | 0.14472 78912 | 0.03418 36735 | 0.00357 14286 | 0.08690 47619 | 0.00527 21088 | 0.00697 27891 | 0.01037 41497 | 0.00357 14286 | 0.00187 07483 | ... |
| 5 | 0.01206 89655 | 0.10977 01149 | 0.11551 72414 | 0.04080 45977 | 0.00057 47126 | 0.06954 02299 | 0.10977 01149 | 0.10977 01149 | 0.00632 18391 | 0.00057 47126 | 0.00057 47126 | ... |

Then to make predictions, what is done needs to take the distribution of words from each topic and later will use Euclidean dot-products [15] to calculate total weight between new documents used for testing with word distribution on each topic. So for example we will predict a new document from the motivation aspect:

*"Saya mencoba semua perusahaan sih Pak yang membuka lowongan."*

Then the data will be preprocessed into :

*['coba', 'semua', 'usaha', 'buka', 'lowong']*

Then we will get every distributions of those words in every topics. We will ignore the word distributions in common topic and we will check in our created model if model does not have the word or term, the weight will be 0 or that word or term will be ignored.

Table 4.42: Word Distributions inside Motivation Model that occurs in new document

| Word or term | KS | K | C | B | BS |
|---|---|---|---|---|---|
| coba | 0.0565072871 3883917 | 0.0144102472 86959615 | 0.0129081245 25436599 | 0.0001203224 6420406689 | 0.0003688675 7654002215, |
| semua | 0.0207108156 48171822 | 0.0001779042 874933286 | 0.0053151100 98709188 | 0.0049332210 32366742 | 0.0003688675 7654002215 |
| usaha | 0.0641779596 0112503 | 0.0873510051 5922434 | 0.1040242976 4616554 | 0.1192395620 2623028 | 0.0704537071 1914423, |
| buka | 0.0283814881 1045768 | 0.0055150329 12293186 | 0.0053151100 98709188 | 0.0013235471 062447359 | 0.0003688675 7654002215 |
| lowong | 0.0437228330 35029406 | 0.0072940757 87226471 | 0.0103771197 16527461 | 0.0145590181 68692092 | 0.0225009221 68941347 |

*After that count the total words occuring in the new document and multiply it with its word distributions.*

*['coba', 'semua', 'usaha', 'buka', 'lowong']*

KS = 1 * 0.05650728713883917 + 1 * 0.020710815648171822 + 1 * 0.06417795960112503 + 1 * 0.02838148811045768 + 1 * 0.043722833035029406 = 0.2135003835336231.

We will do this to every category (K, C, B, and BS).

K = 0.11474826543319694

C = 0.13793976208554798

B = 0.1401756707977379

BS = 0.09406123201770562

Then after everything has been calculated, look for the highest value of the calculation results in each category. If the category has the biggest calculation result, it indicates that the new document is classified as that category. So based on the example above, document *"Saya mencoba semua perusahaan sih Pak yang membuka lowongan."* is classified as KS.

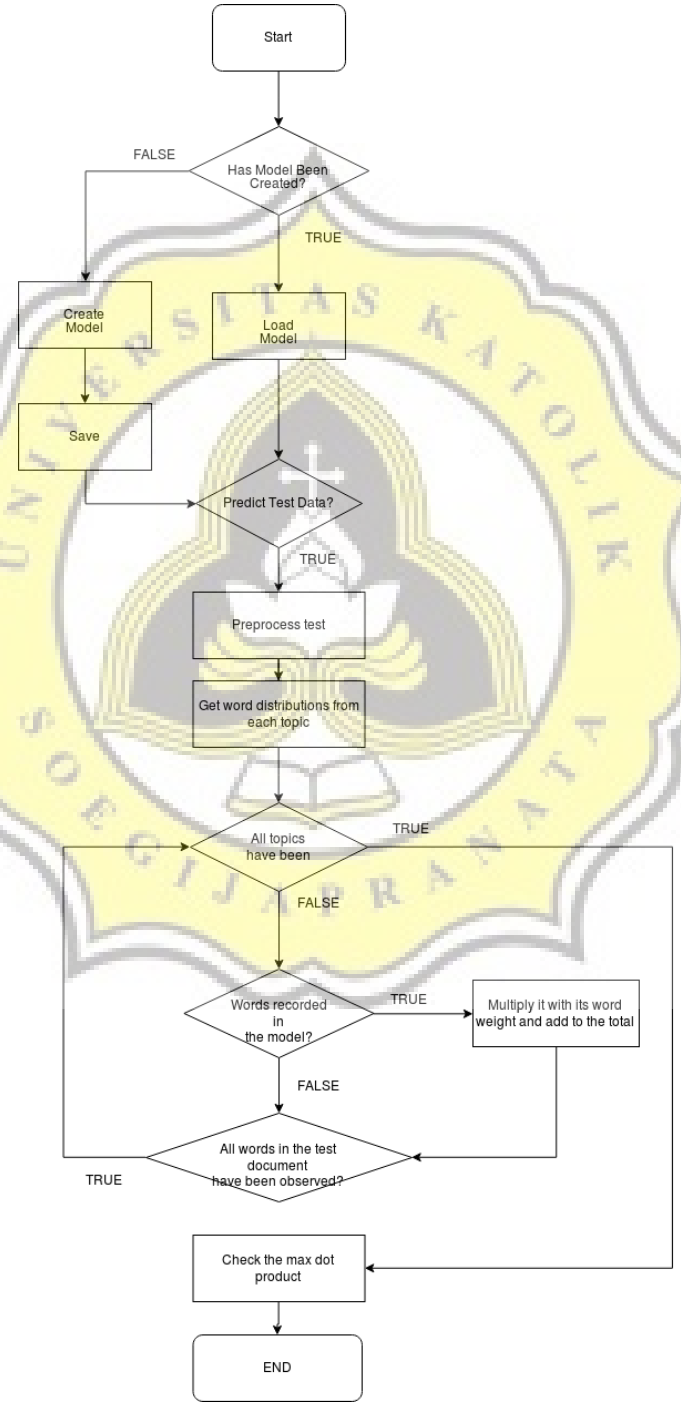
Perform all predictions in the same way to test 100 data in each aspect and then compare the accuracy of the two approaches.

To count accuracy :

$$Accuracy = \frac{The\ total\ correct\ classification}{Total\ testing\ data}$$

## 4.2 Desain

Illustration 4.1: Flowchart



The diagram shows the workflow of the process of classifying each data by aspect. Starting from the initial stage, of course whether the classification model has been made, if not, it will be made first in accordance with the method that has been explained in the analysis section earlier. If it is already created, then the model will only load.

For predictions are also the same as described in the analysis section. Do not forget that the topic here is described as the KS, K, C, B, and BS categories. And we can describe KS, K, C, B, and BS categories into 1, 2, 3, 4, and 5.