Data Understanding

Fase ini merupakan tahapan berikutnya setelah dilakukan pemahaman terhadap tujuan bisnis atau penelitian. Terdapat beberapa proses yang dilakukan dalam tahapan ini, diantaranya adalah tahap pengumpulan serta persiapan data (*preprocessing*) sehingga dapat digunakan sebagai dataset untuk dilakukan pengolahan dengan menggunakan algoritma. Pengumpulan data dilakukan dengan cara *crawling* data menggunakan twitter API, untuk mendapatkan data tweet pada media sosial twitter dengan kata kunci *Moderna,Pfizer,AstraZaneca* dan *Sinovac dalam* periode tertentu. Setelah proses pengumpulan data selesai, dilakukan proses pemahaman data. Proses pemahaman data dilakukan untuk mengetahui kesesuaian data yang didapatkan dengan yang diperlukan serta untuk memberikan label pada data yang digunakan sebagai *train data*, serta untuk menentukan tahapan apa saja yang dilakukan dalam tahapan *pra-processing* dimana salahsatu tahapan di dalamnya yaitu pembersihan data. *Pra-processing* perlu dilakukan untuk menetukan standar dari setiap data yang akan diolah sehingga saat dilakukan pengolahan, dataset memiliki format yang sama serta untuk memperbaiki struktur dari data. Dalam *pra-processing* juga dilakukan tahapan pembersihan data yang bertujuan untuk membuang duplikasi data, menghilangkan atribut yang tidak memiliki makna atau tidak digunakan dalam pengolahan data seperti tanda baca tertentu, url, hashtag (#), mention (@), serta spasi yang berlebihan.

Proses Data Preparation

```
def tweet_cleaner(tweet):

    #remove emoticon and emoji
    tweet = re.sub(r'<U?\+[a-fA-F0-9]\w+\>','', tweet)
    #remove escape caharacter
    tweet = html.unescape(tweet)
    #lower case
    tweet = tweet.lower()
    #remove url
    tweet = re.sub(r'https?:\/\/.*[\r\n]*','',tweet, flags=re.MULTILINE
)
    #replace consecutive non-ASCII characters with a space
    tweet = re.sub(r'[^\x00-
\x7F]+','', tweet).encode('ascii','ignore').decode("utf8")
    #special symbol (hashtag and mention) on tweet
    tweet = re.sub(r"#(\w+)", ' ',tweet, flags=re.MULTILINE)
    tweet = re.sub(r"@(\w+)", ' ',tweet, flags=re.MULTILINE)
    tweet = re.sub(r':','', tweet)
    tweet = re.sub(r',Ä¶','', tweet)
    #remove whitespace and line break
    tweet = tweet.strip()
    tweet = re.sub(r'(?<!\.)\n', '', tweet)
    #remove punctuation
    tweet = "".join([char for char in tweet if char not in str.punctuat
ion])
    tweet = re.sub('[0-9]+','', tweet)
    tweet = re.sub(r'\n','',tweet)
    tweet = re.sub(r'\r','',tweet)
```

```
#remove double word
tweet = re.sub(r'\b(\w+)( \1\b)+', r'\1', tweet)
#split data
tweet = re.split('\W+', tweet)
dic={}

for i in DATA_KBBI:
  (key,val)=i.split('\t')
  dic[key]=val

# kbbi cocokan
tweet = ' '.join(dic.get(word, word) for word in tweet).split()

return tweet
```