

Introduction

Forest fires and haze are very common, it can be said that every year there are forest fires in Indonesia. These forest fires usually occur during the dry season. When the forest is too dry because the water content and humidity are very minimal. The area affected by forest and land fires in 2019 has increased compared to 2018. Based on data from the Ministry of Environment and Forestry of the Republic of Indonesia, the total area affected reached 2.635.223,54 hectares. This number increased compared to 2018 which reached 769.694,97 hectares (Kementerian Lingkungan Hidup dan Kehutanan, n.d.). The impact of forest fires is devastating in many aspects, disrupting health, economy, transportation, and even politics, especially foreign policy, as the smoke from forest fires can spread to neighboring countries, disrupting the lives of people in those neighboring countries (Hakim et al., 2019).

A corpus is a systematically organized collection of natural texts, both spoken and written (Budiwiyanto, n.d.). Although there have been studies about corpus development, there are still few studies that discuss Indonesian and Malay together, a lack of studies that discuss the same vocabulary but have different meanings between the two languages, and manual.

annotation methodologies that have not been widely used for the construction of both corpuses. The purpose of this research is to manually annotate a corpus for Bahasa Indonesia and Bahasa Melayu. A corpus for both languages has been created by many researchers before, but the focus of this research is on words with the same vocabulary but different meanings in the current forest fires. Data was obtained from social media, so informal words were found and then randomly selected with the same vocabulary but different meanings.