

## CHAPTER 1: INTRODUCTION

This chapter presents a general overview of the thesis. Section 1.1 provides the background and Section 1.2 presents problem statement and research question. Sections 1.3 and 1.4 provide objectives and hypothesis. Section 1.5 describes research and methodology of this research and Sections 1.6 lists the thesis overview. Finally, Section 1.7 describes the contribution of this research.

### 1.1 Background

Since 1990s, the internet has been growing so far and becoming a powerful media to gain information from. Some of which is used to obtain the voice of customers by using microblogging channels known as social media. Customers now have enough trusted information from their global peers such that they can select the best product and the best service provider available to them every single time. Every single action on a social network is a proactive expression of what a customer sees, thinks and feels. Smart companies will embrace social support and social selling [1].

This statistic presents the social network penetration in Indonesia. As the fourth quarter 2016, 40% of the total Indonesian population were active with social media users. As seen on Figure 1 the most popular social platforms are YouTube, Facebook, Tweeter and Instagram with more than 50% of penetration rate [3].

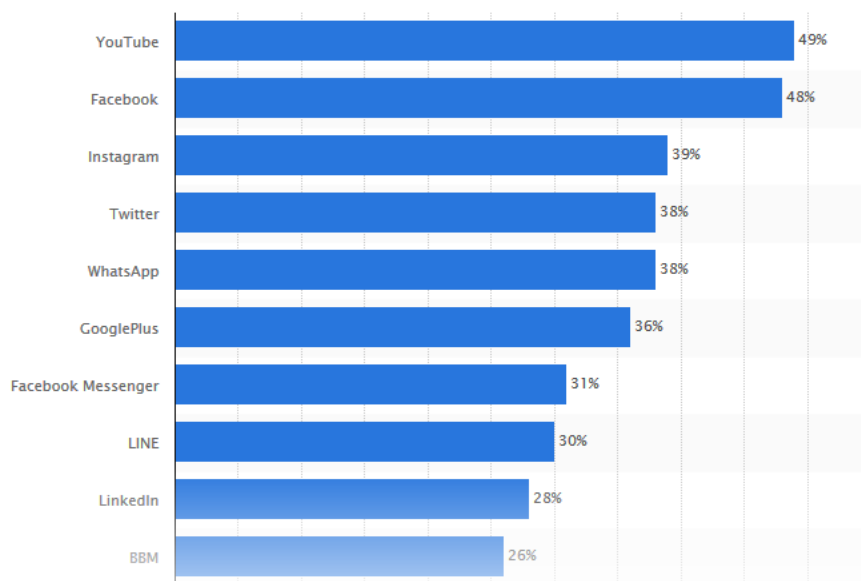


Figure 1. Penetration of leading social networks in Indonesia (4th quarter 2016)

In Telecom industry, providers such as Telkom Indonesia are facing the period where competition leads to creativity, no longer to the price war. The most creative products will win the battle field. The creative products are the products that can fit customer's needs and attract their interest. Creativity should be achieved by looking for new product ideas and perceptions from Telkom product users: the customers.

With the rise of social networking era, there has been a surge of user generated content. Microblogging sites have millions of people sharing their thoughts daily because of its characteristics, short and simple manner of expression. Sentiment analysis is usually used as an integral part of social listening. Sentiments analysis are often being used to analyze the user customer opinions regarding brand images or products [1].

## **1.2 Problem Statement and Research Question**

The main problem in identifying sentiment in sentences, is that sometimes the meaning of opinions is not clearly implied. Despite the high usage of the sentiment analysis application, there are still rooms for improvements. One of the problems that still become a challenge in sentiment analysis is sarcasm, contradictory. Research has been conducted to identify sarcasm in sentences especially in Bahasa Indonesia, where their observation found that in Indonesian social media people tend to criticize something using sarcasm for certain topics [22]. The results of the study found that the additional features for recognizing sarcasm can increase the accuracy of up to 6%, by using negativity and interjection words. In spite of the accuracy did not increase rapidly, this research did provide encouraging results to reduce the inability of computers to recognize sentences that have double meaning.

Another way to solve the inability of computers to recognize those kind of sentences is by proposing human interaction into sentiments identification. Involving human works for online analyzing over large scale data is called crowdsourcing.

Crowdsourcing is an online and distributed problem-solving and production model that has emerged in recent years. Crowdsourcing involves human interaction as volunteers or hired workers to find specific solution of a problem [6]. Notable examples of the model include Wikipedia iStockphoto, Inno-Centive, and Amazon's Mechanical Turk (Mturk) - an online market place for work that requires human intelligence. While the latest in Indonesia was applied on [www.kawalpemilu.org](http://www.kawalpemilu.org), which showed the power of netizen voluntarily providing their effort to validating the final results of Indonesian Presidential Election in 2014. Since the sentiments retrieved from the internet are large-scale data, human involved in this progress should be plenty [15].

Some studies have been done involving crowdsourcing approach. Barbier [16] states that crowds of people can solve some problems faster than individuals or small groups. A crowd can also rapidly generate data about circumstances affecting the crowd itself. This

crowdsourced data can be leveraged to benefit the crowd by providing information or solutions faster than traditional means.

Other research such as Djelassi [17] focuses on customer's participation in a product development process through crowdsourcing practices. Crowdsourcing generates a win-win relationship, creating value for both firms and customers. The results suggest the need to establish an open business model based on crowdsourcing.

On the same way Mukherjee [18] used Amazon Mechanical Turk to crowdsource fake hotel reviews. Truthful reviews were obtained from the TripAdvisor Web site. They attempted several classification approaches which had been used in related tasks such as genre identification, psycholinguistic deception detection, and text classification.

Some other research related to the use of crowdsource to solve sentiment analysis are shown in Table 1.

Table 1. Existing Crowdsourcing Research

<b>References</b>	<b>Research Title</b>	<b>Research Result</b>
Geoffrey Barbier, 2012 [16]	Maximizing Benefits from Crowdsourced Data	Crowdsourced data can be a valuable data source during times of crisis [16]
Anthony Brew, 2010 [19]	Using Crowdsourcing and Active Learning to Track Sentiment in Online Media	Rather than relying on polarity judgments from a single expert, such as an individual economist, the strategy adopted in this system is to generate trend statistics by collecting annotations from a number of non-expert users [19]
Sheng H. Bao, 2014 [20]	Using Crowdsourcing to Improve Sentiment Analytics	Proposing a computerize method of crowdsourced tasks to address computer identified gaps [20]
Mehdi Hosseini, 2014	On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents	Solving the problem of acquiring relevance judgements for information retrieval (IR) test collections through crowdsourcing when no true relevance labels are available.
Arthur Carvalho, 2016 [15]	How Many Crowdsourced Workers Should a Requester Hire	Investigating the optimal number of workers a requester should hire when crowdsourcing tasks, with focus on the crowdsourcing platform of Amazon Mechanical Turk [15]

Telkom itself has started to utilize social media as a great resource for its customers to obtain useful information for the sustainability of their business. One of them is by providing free service on Telkom UData website at <http://socmed.udata.id> (Figure 2). This portal

provides free facilities for public to obtain sentiments based on certain keywords that are very useful to analyze a particular issue related especially to customers.

In this research, the sentiments gathered from UData are used as pre-processing data for crowdsourcing purpose. Referring to machine algorithm, UData itself uses classification for sentences Naive Bayes Classifier to determine sentiments on social media, as stated by its developer. Some improvements are needed because machine learning to improve the accuracy of NLP intelligence especially in Bahasa Indonesia are still about to plan<sup>1</sup>. With current conditions, some tweets containing sarcasm, and contradictions are not completely recognized.

This study is proposing crowdsourcing since UData is a third party built-in app and creating new training dataset is the only way to improve its performance. Another reason for proposing crowdsourcing is in order to find another alternative methods of sentiment analytics other than machine-based algorithm that has been done by many researches.

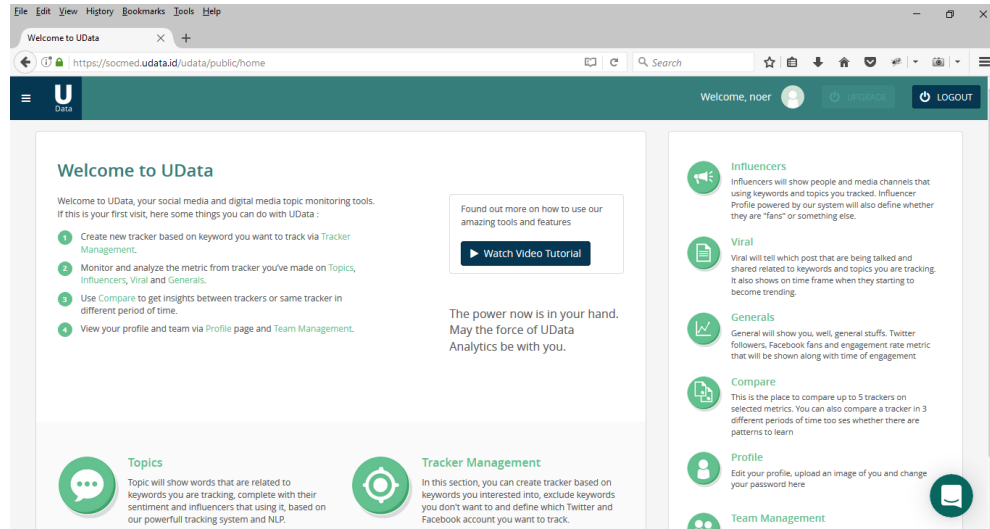


Figure 2. Telkom UData platform at socmed.udata.id

As an interesting example, when the keyword "indihome" is entered then the result of sentiment with the composition is obtained as shown by Figure 3, where the neutral percentage (contains no polarized sentiment) has the greatest composition 46%.

<sup>1</sup> <http://www.udata.id/socmed/#features>

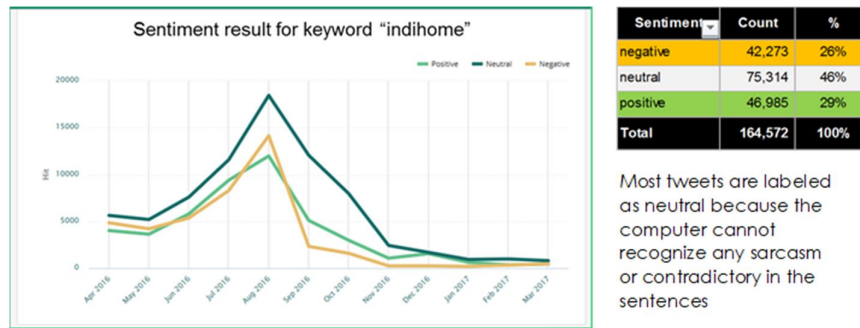


Figure 3. Tweets with keyword 'indihome' retrieved through Apr 2016 to Apr 2017

Then, looking at further particularly for those sentences polarized as neutral, for instance as shown in Figure 4 below, there came curiosity where the sentence 'Bisa mohon dibantu untuk proses berhenti berlangganan indihome via twitter?'. This statement should be negatively polarized, and UData did not recognize it appropriately. With Telkom's point of view as a product owner, this is the sample of losing customer opinion which is not explicitly stated.

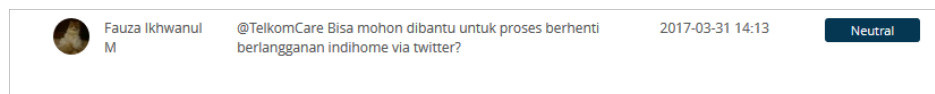


Figure 4. Sentence polarized as neutral

Regarding to the background stated above, it is identified the research question as follows:

1. How can we increase the accuracy information of UData, in this case sentiment analysis obtained from social network (Twitter), by involving human interaction via crowdsourcing approach.
2. How can we identify trust of the crowdsource workers, in order to improve the trustworthiness of sentiments obtained.

### 1.3 Objectives

The objectives of this thesis are:

1. Collecting social media resource to obtain the social perception (sentiments) via Telkom UData, as pre-processing data.
2. Designing a framework for validating the sentiments based on crowdsourcing approach, validated sentiments are used as post-processing crowdsource data.
3. Performing a model to compute trust of the crowdsource workers.
4. Comparing both results from sentiment and crowdsourcing results to check significant improvement.

## **1.4 Hypothesis**

The hypothesis of this research is by adapting crowdsourcing method in constructing the training dataset can improve the sentiment obtained from the machine-based algorithm. Words found in Twitter mostly used are unstructured and informal. While some grammatical structures that form the phrases, clauses, or sentences can be recognized by the machine, there are still many sentiment of sentences that are not well defined. This is mostly due to the uniformity of the informal language used in the social media sentences. Thus human interaction through crowdsourcing method become a proposal in this research.

The crowdsourcing result should produce more accurate sentiment of opinion from customers, so it can provide better input for Telkom management to take action on what to choose for further product development. The result can be useful for monitoring customer needs, analysis of future business growth, predicting the future business trends, as well as more accurate decision making, to get more fresh ideas or to find a new business opportunity.

Moreover, we can also compute trust weight of each crowdsource workers to further strengthen the credence of the sentiment they propose.

## **1.5 Scope and Delimitation**

### **Scope**

This research uses the sentiment retrieved from Telkom UData, a free-to-access social media analytic portal developed by Telkom. The crowdsource workers participated in this research are taken from internal Telkom employee, based on assumption they have well educated about Indihome and familiar to recognize various opinion in social medias.

### **Delimitation**

This research is limited to Telkom product (i.e. indihome), and the crowdworkers involved are only for Telkom employees. In this case all tweets obtained from twitter are filtered for sentences that have only keyword -indihomeø in them, to focus on sentiment of that specific Telkom product.

Since the crowdworkers involved are restricted to within Telkom Group, this crowdsourcing model is known as Enterprise Crowdsourcing.

## 1.6 Research Methodology

The steps of research methodology that were made for the completion of this thesis are as follows.

### a. Problem Identification

This step aimed to identify the potential improvement in sentiment analysis validation by using crowdsourcing mechanism

### b. Model Design

Model design comprised from functional requirements and preparation for design implementation describing how a system was formed. It could be defined as a depiction and making arrangement of several separate elements or sub process into a integrated model. In model design, some suggestions problem solving logic and perform system design problem solving in detail were made.

### c. Data Collection and Processing

Data were collected from data obtained from Telkom UData and convenience sampling. The collected data then ran into data conversion to expected format dataset.

### d. Implementation

The logic problem solving was converted into web application.

### e. Experiment

This experiment aimed to prove the hypothesis in section 1.4 that crowdsourcing result should produce much more accurate sentiment analysis. Some scenarios were performed to find out the experiment results.

### f. Testing Result and Analysis

This section calculated the accuracy of the proposed system and comparing with previous system. The analysis of experiment results was performed.

## 1.7 Thesis Overview

This thesis is organized in five chapters consists of introduction, literature review, algorithm design and implementation, experiment and data analysis and conclusions. The explanation each chapter as follows:

### a. Introduction

This chapter discusses the background of the issues raised in this thesis, the problem that becomes reason of this final project, the research question, objectives reached, research methodology at glance and thesis overview.

### b. Literature Review

This chapter discusses the theoretical basis of the concepts and theories that support and used in this final project. Literature review refers to the theoretical basis that is used as a base reference guide to the final project.

c. Research Methodology

This chapter describes the general description of the system and the proposed model for completion existing problems in this thesis. In this design it describes general description of the system both in terms of the structure of the system in detail, the proposed model, data flow, as well as system usage scheme.

d. Experiment and Analysis

This chapter describes the purpose of testing and testing scenario for completion existing problems in this thesis. At the stage of the testing results analysis, it explains about data that coming from testing results, the graph of the test results and its analysis.

e. Conclusions

This chapter explains the final conclusion of this study. Conclusions include the results of the final analysis that has been done and explain the answer to the problem statement that formulated in this thesis. At the recommendation point, it describes suggestions regarding the possibilities that can be developed from the results of this research.

## 1.8 Contribution

This thesis has two primary research contributions:

- a. Crowdsourced-sentiment method of validating machine-generated sentiment analysis by using crowdsource platform, with a measurement of trust level between crowdworkers to increase the confidence of sentiments they offered.
- b. Dataset used in this thesis still can be developed for the future research, especially to improve and enrich Telkom UData sentiments used as a case study in this thesis.