CHAPTER 1

INTRODUCTION

Twitter is one of the largest social media with about a billion unique visit per month as 30 September 2015. This situation gives advantages to its users to share information rapidly within a short time. There are some functions of Twitter where users could interact with each other; *trending topics*, *private message*, *time line*, *mention* and *re tweet*. The most popular part that is visited by many people who use Twitter is *trending topics*.

Twitter *trending topics* is a function on Twitter where tweets related to topic matters around certain region are collected around some hash tags. Hash tag is a keyword followed by hash symbol. By typing the hash tag e.g. *#papamintasaham*, *#SBMPTN2016*, tweets around that topic can be accessed. This feature is the easiest method to make an information exposed to many people at once. Every tweets that has hash tag will be seen by every people who use same *hash tag*.

Unfortunately, the very same principle is used by some people to send unrelated messages to its *hash tag*. On Twitter policy, Twitter stated that "posting repeatedly to *trending topics* to try to grab attention" is one of their description of Spam[12]. What is more interesting about the trending topic is it is region specific. It is very alluring for people who want to Spam inside some target specific area. By inserting *hash tag* which is popular in certain area, people could easily expose their messages to many people in certain region. Indonesia is one of five top user in Twitter [13] and it is has been known that around 9.3% Twitter consists as Spam[14]. Based on that facts, Indonesian who uses Twitter will be most likely get exposed more to Spam.

Spam has become one of important problems for Twitter to deal with. Through Spam, users of Twitter could be exposed in danger of malicious activity, such as like-jacking, fake accounts, Spam applications to phishing activity[13] that could lead to on line identity theft. Therefore, reducing Spam in Twitter is important for safer and cleaner Twitter as social media.

The background of this research is discussed on section 1.1. The concept and overall aspects that will affect the research are presented in section 1.2. Section 1.3 explain about problem that this research tried to solve. Section 1.4 explains aim that this research want to achieve. While section 1.5 describes the hypothesis on some aspects on this research. In Section 1.6 scope and delimitation of this research is discussed. Contribution of this research is stated on section 1.7.

1.1 Rationale

Research in Spam detection always use human tagged data in order to build a Spam filter. It has been known that human tagged data is highly preferred because human could adapt to unknown condition when gives label to a data [15]. However, this method is quite expensive as it needs a lot of human involvement. Every human who work inside the system should be compensated in form of money to entice its worker [16]. To address this problem, there are some studies to create a way to extract high quality human tagged data at affordable cost, e.g MTurk [17]. However, this method is still very costly in order for the system to run properly.

Detecting Spam in general languages is very hard, this includes in Indonesian language. According to KBBI, Indonesia have approximately 92.000 official words. Furthermore, language that is used in social media is usually diverted from the official language. Word limitation on Twitter also encourages people to use many slang abbreviations. To correctly determine a Spam, collection of words is not sufficient, the context when and how those word are used and connected is very important. Context recognition is a basic trait of human being. So, it should be more effective to directly use human ability to recognize context intuitively rather than what computers do with some procedures.

Human computation is a way to utilize human thinking power and their trait to adapt for the unknown as a processor in a distributed system [18]. This system has been successful in many computation problems which are very hard tasks for computers but rather easy for human[17–26]. There are many similar terms related closely with human computations[27], such as crowd sourcing, social computing, data mining and collective intelligence. In their research, von Ahn and Dabbish[22] tried to tag web image using Game With a Purpose, the result is high quality web images label that is 85% useful to describe the image.

Problem that von Ahn and Dabbish[22] solved in their research was vision or image understanding, an AI-Complete problem. Another variant of AI-Complete problem is Problem Solving, Knowledge Representation and Reasoning, and Natural Language

Understanding. AI-Complete problem is a computer problem that could be solved only if general artificial intelligence has been established[28].

Currently, to solve this problem, trait of AI-Complete problem was utilized, which is a problem that is human oracle solvable. Therefore, human computation is suitable to solve AI complete problem.

Spam is a problem that is categorized into natural language understanding problem, which is in the same set as image understanding that has been solved by von Ahn and Dabbish[22] which is AI-complete problem. Therefore, by using human computation, solve Spam problem in Indonesian Twitter trending topic hopefully could be solved.

1.2 Conceptual Framework

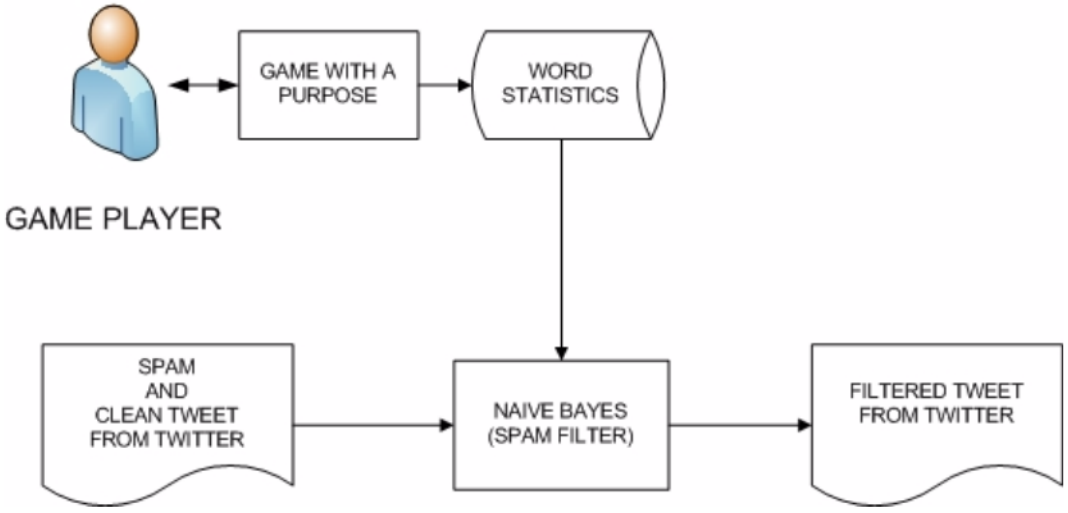

FIGURE 1.1: Conceptual Framework

The concept of our method could be seen on figure 1.1, the process started when people give their input to game with a purpose system. Product of this process is word statistics that will be used by Naive Bayes to filter Spam from Spam and clean tweet from Twitter, the end result will be filtered tweet that should be cleaner than before.

1.3 Problem Statements

Spam is an AI complete and a recurring problem that still happens in Social media. There are previous research that tried to solve this problem[31][32][33][34]. Despite their effort there is still some Spam that is visible in Twitter trending topic, implying that Spam problem still not yet solved by these research.

1.4 Objective

This research aim to detect Spam that is lingers on Twitter in Indonesian Trending Topics . In order to reach the aim, we will build a Game With a Purpose system to train Naive Bayes that will be used as Spam filter.

As a prerequisite before building the system, we will search for respondents and do some research on how develop the game. After the game has been build, we will conduct the experiment where the respondent plays and then the Spam filter will be optimized. Data from the experiment then will be analyzed in order to conclude the result of this thesis. We also will compare our Spam classifier performance with Yadav et al.[11] Spam classifier in order to see the improvement of our method.

1.5 Hypothesis

As said by [28]Yampolskiy, human computation is known to solve AI complete problem theoretically. It also has been used to solve several of AI complete problem, but not yet utilized to solve Spam problem. GWAP is a form of human computation which take game system as its main data input.

GWAP has been used in order to solve AI complete problem before. In language understanding, sub part problem of AI complete problem, there are duolingo, a GWAP system that have goal to translate the web. The translation result is distinguishable from human like translation [35] which is a sign that it is very good in its performance. Spam problem is similar with translation problem in regards that it is also a language understanding problem. Therefore, the conclusion is that usage of GWAP in Spam filtering problem will yield good result.

However, comparing one GWAP system with another GWAP system is difficult because the problem that is solved on each system is clearly different with each other. Instead, GWAP will be used as means to gather data, and then Spam filter that is generated from the data will be compared with another similar Spam filtering method performance. The plan in this study to detect Spam is closer to Naive Bayes, therefore Spam filter performance will be compared with it.

Naive Bayes is a machine learning method that has been used to filter Spam in Twitter in some research [34][36][37][11]. All of them predicts Spam in Twitter using prior knowledge of previous known Spam. However, as stated by Schneider[38], prior knowledge is not suitable to be used in short document detection because it makes the result

of prediction worst, the example of this document is tweet in Twitter. Therefore, this research believe that by using GWAP based Spam filter, performance that could be achieved by basic Naive Bayes alone will be surpassed.

1.6 Scope and Delimitation

1.6.1 Scope

The respondent that have participated in this research will be taken from Telkom University. The respondents are knowledgeable in social media, especially Twitter. Another assumption is the data that will be fetched from Twitter trending topics is easily accessed through R software. We also assume that people who will become the respondent of this research have adequate visual recognition and language understanding skills.

1.6.2 Delimitation

The area of interest in this research is Twitter trending topic in Indonesian region and language. From that section, data from top hash tag in Twitter trending topic was taken. The hash tag that was used in this research is #BigSupportToRaffiAhmad, a hash tag taken from 16 November 2015 and #SBMPTN2016 that was taken from 31 May 2016. Definition of Spam that is used in this research is a definition that is used by Twitter in their policy.

1.7 Importance of The Study

This research enhances understanding of game with a purpose as a mean to create Spam filter. Furthermore, it will increase Twitter accuracy of Spam filtering of both bot and human created Spam, especially in Twitter trending topic. Finally, The current research will give addition to a growing body of literature on Human Computation.