Chapter 3
Description and Initial Analysis of Cyberbullying Dataset

ABSTRACT

In this chapter, the authors focus on datasets used in cyberbullying detection research. They describe and compare several datasets applied in previous research and describe in detail the dataset that they decided to apply in their research. They also perform an initial analysis of the dataset to find various characteristics. They preprocess the dataset in several ways for further use and perform affect analysis to find out whether emotion-related features tend to be characteristic for cyberbullying. Based on the results of affect analysis, they also perform an initial attempt to classify cyberbullying data using a simple machine learning approach, which will be considered as a baseline in forthcoming chapters.

DESCRIPTION OF AVAILABLE CYBERBULLYING DATASETS

The research on cyberbullying detection started globally around the year 2009/2010. Since then a number of research teams attempted to tackle the problem. The general description of some of the research done previously has been presented in previous chapter. Here, we focus on datasets used in those research. There were several conditions we applied to choose the dataset for description in this chapter. Firstly, the dataset needed to be applied in more than one research paper. Moreover, we focused on datasets which were

DOI: 10.4018/978-1-5225-5249-9.ch003

significantly large, meaning, several thousands of samples or larger, desirably with balanced distribution of samples (cyberbullying to non-cyberbullying). All analyzed datasets were summarized in Table 1.

Historically the first dataset containing cyberbullying messages was the one created originally in 2008 and applied by Matsuba et al. (2009), and since then has been widely used by others (Matsuba et al., 2010, 2011; Ptaszynski et al., 2010; Nitta et al., 2013; Ptaszynski et al., 2015a, 2015b; Ptaszynski et al., 2016). It contains 1,490 harmful and 1,508 non-harmful entries in Japanese collected from unofficial school Web sites and fora, usually represented in the form of Electronic Bulletin Board Systems (BBS). The original data was provided by the Human Rights Research Institute Against All Forms for Discrimination and Racism in Mie Prefecture, Japan (http://www.pref.mie.lg.jp/jinkenc/hp/). The harmful and non-harmful sentences were collected and manually labeled by Internet Patrol members (expert annotators) according to instructions included in the governmental manual for dealing with cyberbullying (Ministry of Education, Culture, Sports, Science and Technology, 2008). More detailed descriptions of the instructions were explained in later sections of this chapter.

In 2009, Xu and Zhu (2010) collected a set of 11,670 comments in English from YouTube videos. They extracted the dataset half automatically, by using a set of offensive words. Because of the way the dataset was collected, it cannot be considered as fully cyberbullying-oriented, since offensive words can appear in a large variety of contexts. However, despite being largely imbalanced (harmful information was less than 15%), the authors later proved, that the corpus can be applied in a task related to cyberbullying detection with some success (Chen et al., 2012).

Also in 2009 a workshop called Content Analysis for the WEB 2.0 (CAW2.0) collocated with The 18th International World Wide Web Conference (WWW2009) released a dataset containing data from a number of SNS (Kongregate, Slashdot, MySpace, etc.) with application in a series of tasks (all data in English). One of the tasks was called "Misbehavior Detection" and considered detecting a generally defined malicious messages posted on such SNS. The dataset was applied in a number of studies on detection of Internet harassment (Yin et al., 2009; Bayzick, Kontostathis & Edwards, 2011).

In 2010, Ishisaka and colleagues automatically collected a dataset of, what they called "nasty comments" from a Japanese BBS forum 2channel (https://www.2ch.net/, recently renamed to 5channel, https://5ch.net/) (Ishisaka &