Towards preprocessing on criminal chatting corpus.

Marjuni, Siti Hanom and Mahmod, Ramlan and Abd Ghani, Abdul Azim and Md Zain, Abdullah and Sidi, Fatimah (2009) *Towards preprocessing on criminal chatting corpus.* MASAUM Journal of Basic and Applied Sciences, 1 (3). pp. 401-405. ISSN 2076-0841

Abstract

The importance of data cleansing is apparent with the advent meaning of the words collected during a conversation. In addition, noise, concise expressions and dynamic situation makes chat data ill-suited for analysis. Due to its often informal nature especially in short form language, this paper will present the importance of preprocessing steps of data collection before we proceed to the next stage of the research. Two processes of cleaning data are required in this research. First, the conversion of short form words to full English words and second, discarding all toggles found in every utterance of the conversation. The processing is to make the sentence more meaningful due to the suspect's target and expectation of intention. Results done by precisions, recalls and f_measure showed that the corpus need the conversion to be more meaningful. Furthermore, each word of the suspect's and victim's utterance is analyzed and treated as support evidence in criminal court cases. This research will consider criminal data chatting through Yahoo Messenger (YM) which involved the suspect's and victim's conversation collected in real time without any editorial changes in electronic discourse. However, chat messengers are in an unstructured format which always use short form languages. Chatters may use the typical language or use their own understood language during the conversation. Therefore, we propose the preprocessing phase for specifically chat data mining which involve text messages. The idea of the preprocessing is to prepare cleaned data called corpus criminal data and the cleaned data will be used in the next phase for identifying words classification, tokenizing, tagging, ranking and constructing the meanings.

Item Type: Article

Keyword: Criminal chatting; Data cleansing; Preprocessing; Short word.