

Abstract—Social media is a forum where individuals congregate and exchange information, thereby facilitating contemporary communication. Social media content presents unique challenges for information extraction due to its diverse and unstructured nature. For instance, the most prevalent content on social media platforms such as Twitter/X is commercial in nature, typically consisting of product advertisements or product promotions. The primary obstacle to accepting the information is the diverse and variable structure of the user's writing styles, which present a variety of formats. The multitude of writing formats within the information in question renders it less efficient. This is where the role of information extraction comes in, transforming the unstructured information into a structured format using the Bidirectional Long-Short Term Memory with Conditional Random Fields (BiLSTM-CRF) method. This method was selected because it provides context for information from both the past and the future, which is suitable for the task of extracting information. The objective of this study is to extract information using a Bidirectional Long Short-Term Memory (BiLSTM) with Conditional Random Fields (CRF) for classification. This process involves the extraction of information in accordance with the BIO labels, which are then structured and made accessible for analysis. The results obtained from the implementation of the aforementioned model yielded the following performance values: precision of 87%, recall of 77%, and an F1-score of 80%.

Keywords-social media, product advertisements, BiLSTM-CRF, information extraction