Information Extraction of Product Advertisement Post in Social Media Using BiLSTM-CRF

***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

I. Introduction

A. Background

This study investigates the implementation of BiLSTM-CRF architecture for processing online shopping data from X platform, a popular social media platform where users actively share information and experiences related to online shopping. Through tweets, images, and videos, consumers frequently share product reviews, recommendations, criticisms, and their online shopping moments. This information is invaluable for e-commerce companies, marketers, as well as other consumers in understanding market trends, consumer preferences, and product quality [12]. However, extracting relevant information from X's unstructured, informal, and often noise-containing data such as abbreviations, slang, and typos is a challenge. Traditional Information Extraction (IE) methods are often not effective enough to handle the unique characteristics of social media data such as X [4]. Recent advances in deep learning have yielded significant improvements in Information Extraction tasks, particularly in Named Entity Recognition (NER) as a fundamental component of information extraction [6]. The integration of Bidirectional Long Short-Term Memory (BiLSTM) with Conditional Random Fields (CRF) has demonstrated exceptional performance in sequence labeling tasks, especially for NER applications [5]. However, existing research predominantly focuses on traditional domains like news articles and biomedical texts [1]. The application of BiLSTM-CRF models for extracting information from online shopping conversations on social media platforms like X remains relatively unexplored [8]. The distinctive features of X's shopping-related content, including colloquial expressions, abbreviated language, character constraints, and contextual complexity, pose significant challenges for information extraction systems [7]. This research investigates the implementation of BiLSTM-CRF architecture for processing online shopping data from X platform. Through the combination of BiLSTM's contextual modeling capabilities and CRF's structured label prediction, we aim to address the specific challenges presented by social media content [6]. This study's outcomes will contribute to advancing information extraction techniques for online shopping domains on X while enhancing our understanding of effective methods for processing social media content [4].

B. Topics and Limitations

Based on the description in the background, the main problem can be raised in this study, namely how to implement the online shopping information extraction model on X users' tweets using BiLSTM-CRF. The limitations of the issues raised are:

1) The dataset used is online shopping information obtained from X users tweets.
2) Data in Indonesian.
3) The data owned uses the BIO schema format where the labels are divided into:
   - Product: Products offered.
   - Brand: Brand of the offered product.
   - Specifications: Specifications of the offered product.
   - Price: Price of the offered products.
   - Warranty: Availability of warranty for the products offered.
   - Promo: Promos available for the products offered.
   - Online Shop: The store where the product is offered, provided as an online store name or link.

C. Objectives

The purpose of this study is to implement a model that is capable of extracting information on post advertisement activities using the BiLSTM-CRF architecture on X datasets. The results obtained are then analyzed and evaluated on the performance of the implemented model's performance.

D. Writing Organization

This study is structured into five distinct sections. The preliminary section encompasses a comprehensive introduction, which delineates the background, scope of investigation, inherent limitations, and research objectives. The second section presents a critical review of relevant literature and contemporary research pertaining to the subject matter. The third section elucidates the methodological framework, specifically focusing on the implementation of the BiLSTM-CRF architecture and the developed model's operational workflow. The fourth section provides a detailed analysis and evaluation of the model's performance metrics and outputs. The fifth section synthesizes the key findings and implications derived from this study implementation.

II. Literature Review

A. Information Extraction

Information Extraction (IE) is a field in Natural Language Processing (NLP) that focuses on extracting structured information from unstructured text data [3]. The primary objective of IE is to identify named entities, relationships between entities, and events within the text. Key tasks within IE include Named Entity Recognition (NER), Relation Extraction, and Event Extraction [4].

B. Named Entity Recognition (NER)

NER is the task of identifying and classifying named entities, such as product names, brand names, locations, and organizations, within text [8]. NER is a crucial component in many NLP applications, including Information Extraction, Question Answering, and Information Retrieval. Various approaches have been employed for NER, including rule-based, machine learning, and deep learning techniques.

C. Information Extraction of Product Advertisement Post

Information Extraction (IE) of product advertisements from social media presents unique challenges due to non-standard language elements, requiring specialized approaches beyond traditional extraction methods. Social media platforms, particularly X, have transformed how users share retail experiences and product information. X has established itself as a key platform where consumers actively share shopping experiences, product opinions, and retail interactions [1]. However, processing this user-generated content presents unique extraction challenges, as posts frequently contain non-standard language elements like abbreviated text, informal expressions, and typographical variations [2]. Standard Information Extraction (IE) methods face additional hurdles when processing product-related social media content. Posts typically mix casual language with product details, creating dense but fragmented information streams [11]. The content differs significantly from traditional knowledge bases - instead of standardized entity references, social media features diverse mentions of local sellers and user-specific product descriptions that require specialized entity resolution approaches [3]. Additionally, the dynamic nature of social media content introduces data quality challenges, from missing context to contradictory details, requiring robust methods for accurate information extraction under uncertainty [3].

D. Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional LSTM (BiLSTM) is a variant of the LSTM architecture that can utilize context information from forward and backward directions in a sequence [5]. BiLSTM has been widely used for sequence labeling tasks such as NER and has improved performance compared to regular LSTM models.

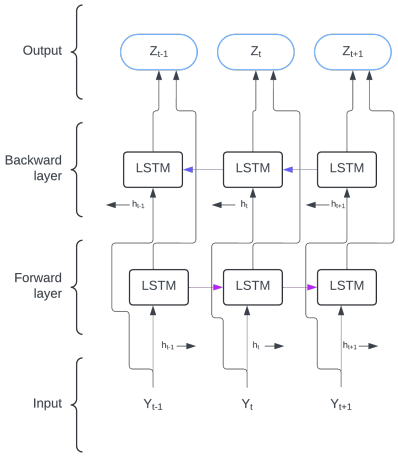

Fig. 1. Example of BiLSTM architecture

E. Conditional Random Fields (CRF)

Conditional Random Fields (CRF) is a statistical model often used for structured prediction in NLP tasks [5]. CRF can take advantage of observations to predict appropriate labels, as well as model relationships between labels in sequence. For example, there is a fruit of sequential observation, for sequence input, where $Xnx = (x1, x2, ..., xt, ..., xn)t(1)$ is the input of the vector t-th word, represents the generic label for $x, y = (y1, y2, ..., yt, ..., yn)(2)$ [9]

Thee combination of BiLSTM and CRF (BiLSTM-CRF) has become a popular model, efficient, and achieves state-of-the-art performance for sequence labeling tasks.

F. Evaluation Method

Modeling requires measurements that will be used to calculate the quality of the model that has been built. Research this research uses common evaluation methods used in classification algorithms, including F1-score, precision, and recall.

Precision prioritizes accuracy over sheer volume of results. A high-precision system focuses on returning only the most relevant items, even if it means discarding some possibilities. This ensures a greater proportion of truly accurate results, at the expense of potentially missing some relevant information [10].

$$Precision = \frac{|True\,Positives|}{|True\,Positives| + |False\,Positives|} \quad (3)$$