Data Corruption Impact on Named Entity Recognition for Low Resourced Languages

Anonymous
Anonymous
anonymous

Abstract

1 Introduction

Natural Language Processing (NLP) is an impactful field, and has received much interest recently. However, much of the focus is on high-resourced languages, such as English, German, Spanish, etc. While this has led to impressive results for these languages, lower-resourced languages have often not received as much attention, leading to worse performance.

While there has been interest in low-resourced NLP, one limiting factor is data. (e.g. cite kreutzer 2022 paper). In particular, much of the data available online has questionable quality. This has led to many works focusing on filtering data (e.g. idris paper WMT) to improve results. We instead take a different perspective, and focus on thoroughly analysing the

References