WORD EMBEDDING AND ITS APPLICATIONS FOR TELUGU LANGUAGE

MRS. G. SANTHOSHI
Assistant. Professor
Department of Information Technology
G Narayanamma Institute of Technology & Science
Shaikpet, Hyderabad, Telangana, IN

WORD EMBEDDING AND ITS APPLICATIONS FOR TELUGU LANGUAGE

Copyright© : Mrs. G. Santhoshi
Publishing Rightsⓟ : VSRD Academic Publishing
A Division of Visual Soft India Pvt. Ltd.

ISBN-13: 978-93-91462-86-4
FIRST EDITION, JULY 2023, INDIA

Printed & Published by:
VSRD Academic Publishing
(*A Division of Visual Soft India Pvt. Ltd.*)

Disclaimer: The author(s) / Editor(s) are solely responsible for the contents compiled in this book. The publishers or its staff do not take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are requested to communicate such errors to the Author(s) or Editor(s) or Publishers to avoid discrepancies in future.

Printed & Bound in India

VSRD ACADEMIC PUBLISHING
A Division of Visual Soft India Pvt. Ltd.

REGISTERED OFFICE
154, Tezab mill Campus, Anwarganj, KANPUR–208003 (UP) (IN)
Mb:9899936803, Web: www.vsrdpublishing.com, Email: vsrdpublishing@gmail.com

MARKETING OFFICE
340, FF, Adarsh Nagar, Oshiwara, Andheri(W), MUMBAI–400053 (MH) (IN)
Mb:9956127040, Web: www.vsrdpublishing.com, Email: vsrdpublishing@gmail.com

PREFACE

Word embedding methods are used to represent words in a numerical way. The machine learning or deep learning algorithms process the text data. The machines cannot understand the text data, it understands only numbers so we need to convert the text data to numerical form by using word embeddings techniques. Representing a word using vocabulary. Next map vocabulary to vectors. One-hot vectors are a quick andeasy way to represent words as vectors of real-valued numbers. In perspective of technology, by using the word embeddings are represented in syntactic form only. Whereas by using the predictive based embeddings, the words are represented in semantic form. In perspective of languages, the Indian subcontinent consists of a number of separate linguistic communities each of which share a common language and culture 22 major languages have been given constitutional recognition. We want to try this technology in the Telugu language. we want to build and test the different word embedding model on different machine learning(ML) algorithms. We tested two embedding model on 3 Machine learning algorithms for Telugu language. The main challenge is with the aggloramatives and inflections. In Telugu there are so many inflections. Telugu is a highly inflected as well as morphologically rich language. A slight modification in a word can change its form to express a completely different meaning from the original one. Using the application data set by comparing the two embedding models with Machine learning algorithms Word2vector model works well compared to One hot encoding model.

CONTENTS