Development and Evaluation of a Tool for the Retrieval of Semantically Related Words for Lithuanian Language

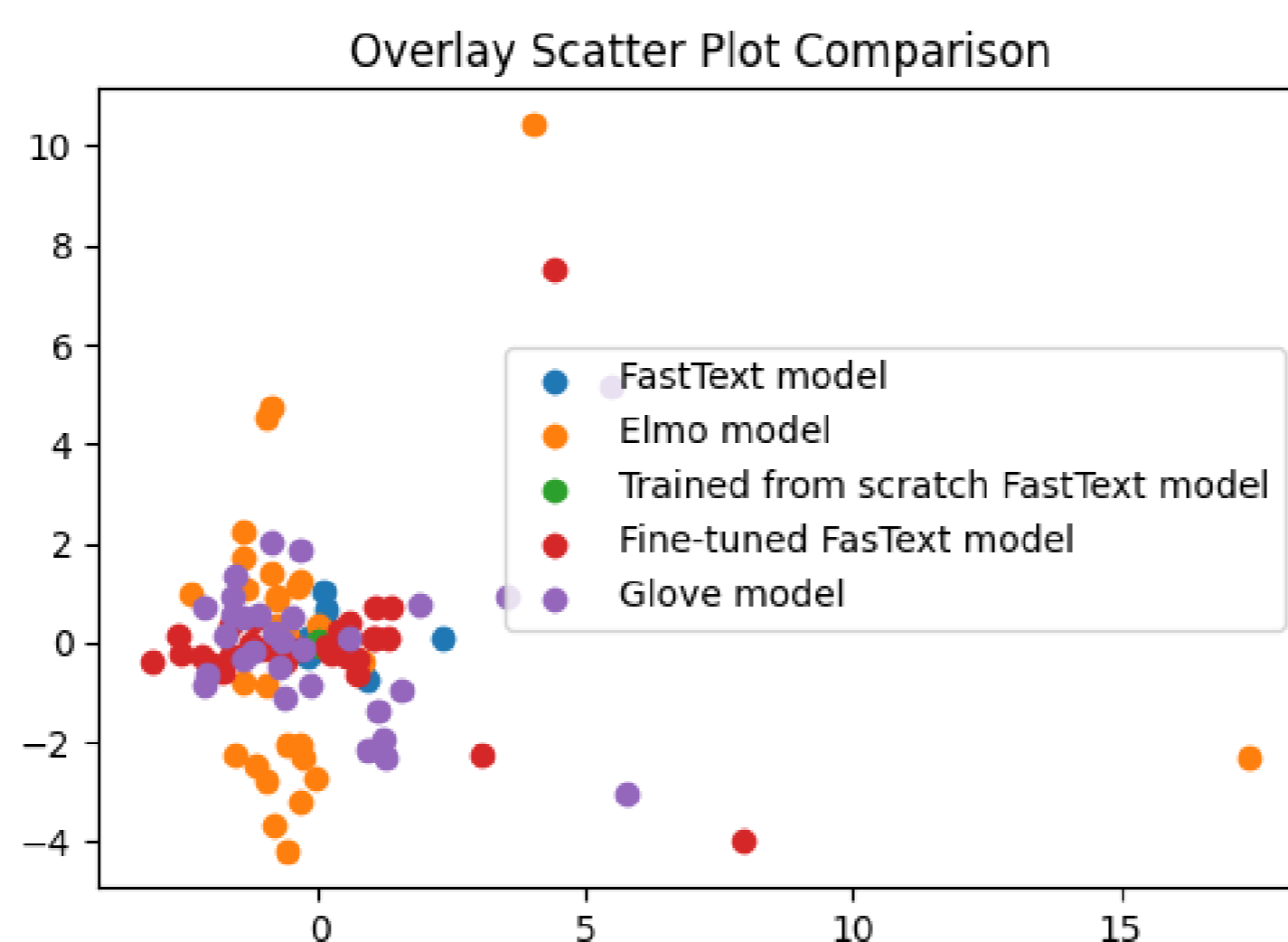VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

INTRODUCTION

- Addresses the lack of user-friendly tools for less studied languages.
- Features a Graphical User Interface (GUI) and request-response mode for easy interaction and semantic exploration.
- Experimented with various word embedding models using the Pedagogic Corpus of Lithuanian.
- Focuses on linguistic precision and depth for a wide range of applications.
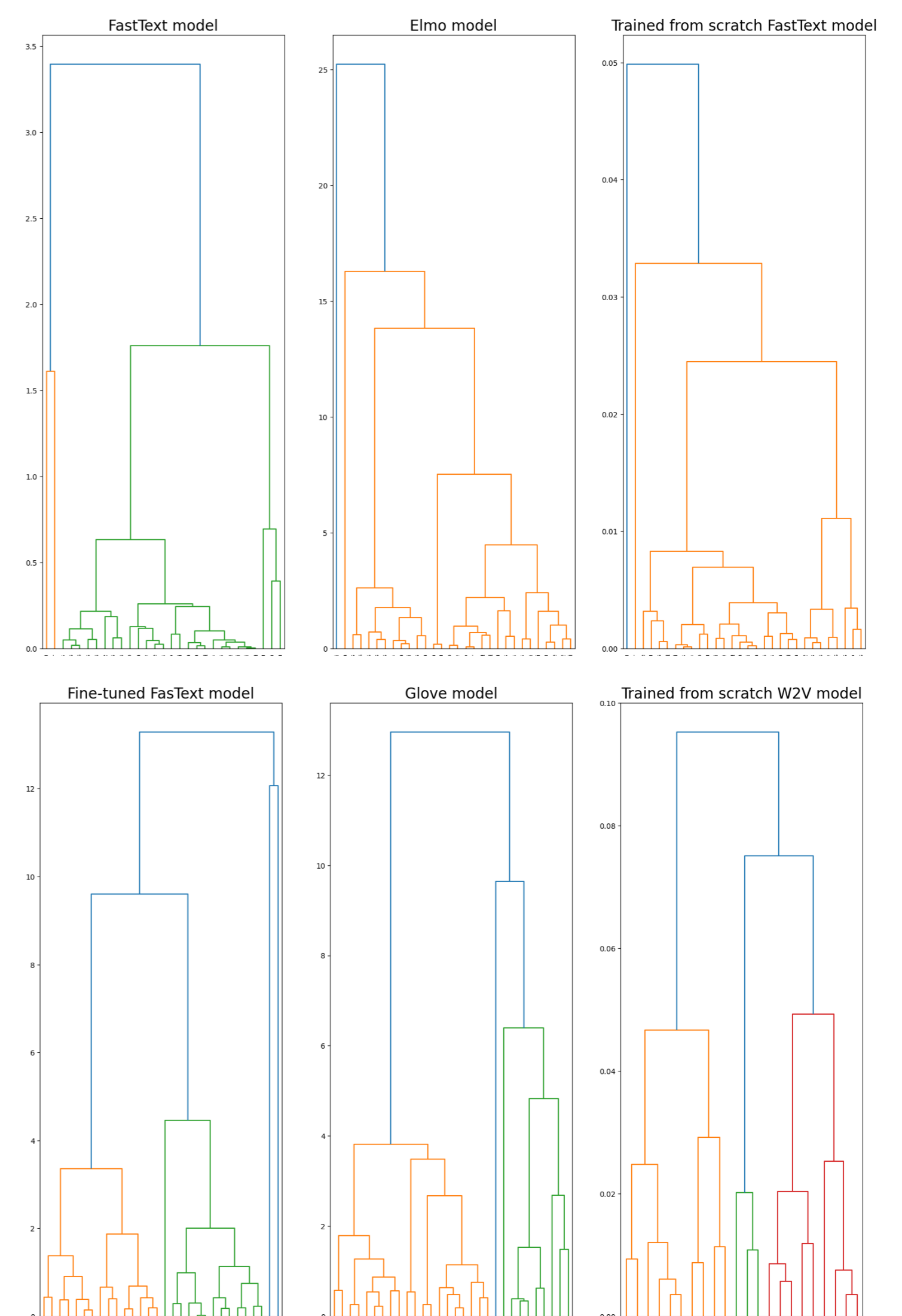
GOAL

- Develop a tool for retrieving nearest word approximations using word embeddings for the Lithuanian language.
- Conduct thorough evaluation and benchmarking of the developed tool's performance
- Facilitate linguistic research and technology development projects
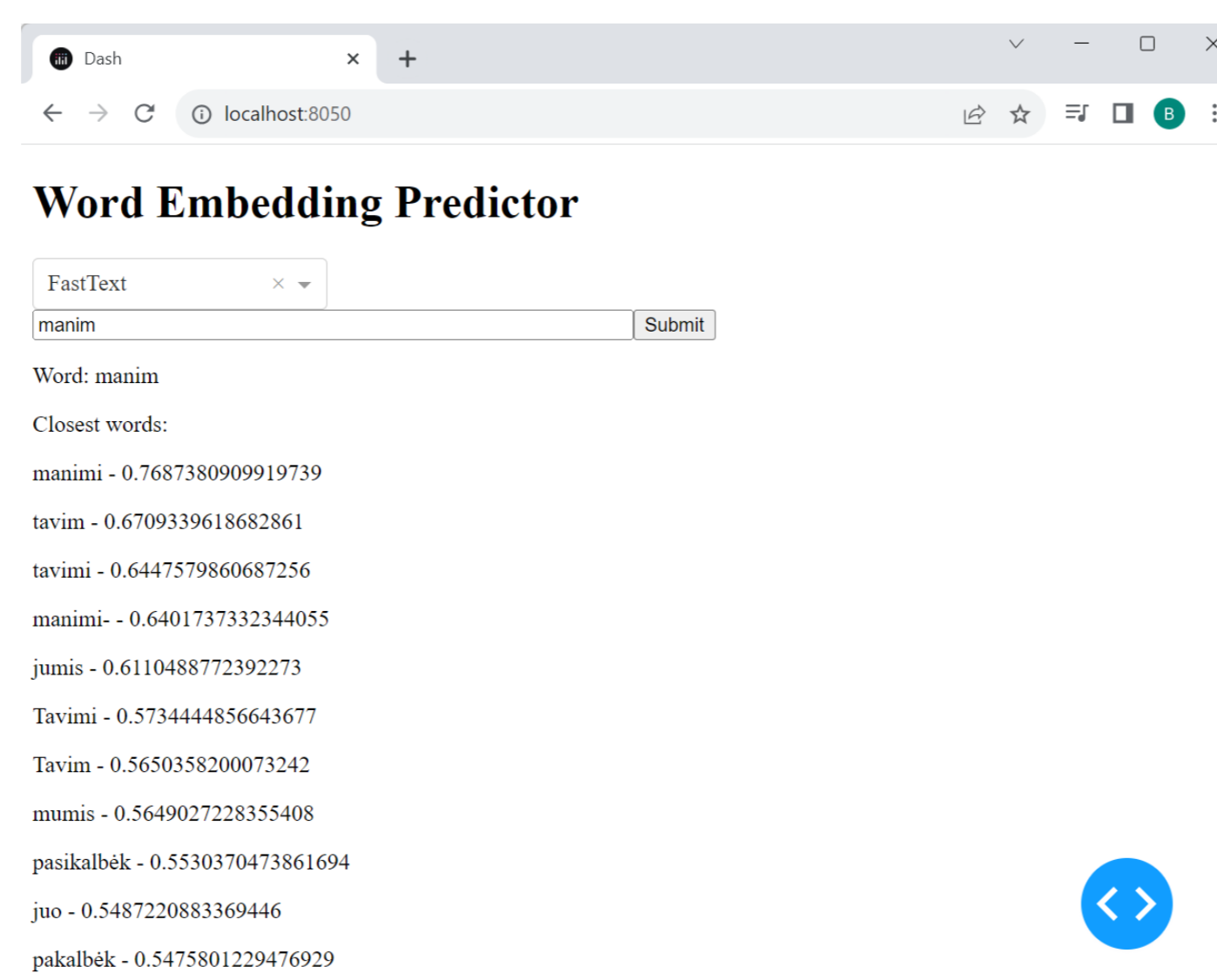.

COMPARISSON

Overlay Scatter Plot Comparison

Legend:
- FastText model
- Elmo model
- Trained from scratch FastText model
- Fine-tuned FasText model
- Glove model

EVALUATION

FastText model | Elmo model | Trained from scratch FastText model

Fine-tuned FasText model | Glove model | Trained from scratch W2V model

RESULTS

Word Embedding Predictor

FastText

manim   Submit

Word: manim

Closest words:

manimi - 0.7687380909919739
tavim - 0.6709339618682861
tavimi - 0.6447579860687256
manimi- - 0.6401737332344055
jumis - 0.6110488772392273
Tavimi - 0.5734444856643677
Tavim - 0.5650358200073242
mumis - 0.5649027228355408
pasikalbėk - 0.5530370473861694
juo - 0.5487220883369446
pakalbėk - 0.5475801229476929

CONCLUSIONS

- Successfully created a unique linguistic tool for Lithuanian.
- Offers user-friendly functionality for word embedding-based queries.
- Provides insights into the performance of different word embedding models.

AUTHORS:

Bohdan Zhyhun
bohdan.zhyhun@vdu.lt

Justina Mandravickaitė
justina.mandravickaite@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT

ACKNOWLEDGMENT

FUTURE PLANS

- Refine and optimize the tool's performance based on user feedback.
- Expand the training corpus to enhance the tool's accuracy and coverage.
- Collaborate with linguistics and technology communities for ongoing improvements.