DIACHRONIC TEXT CLASSIFICATION

A DEEP LEARNING APPROACH TO IDENTIFYING HISTORICAL ERAS IN URDU LITERATURE

AIZA IMRAN
SYED HAMZA

CS/CE 316/365

زبان بنائی نہیں جاتی، خود بنتی ہے

QUOTE BY SAADAT HASAN MANTO [1]

[1] REKHTA FOUNDATION. REKHTA, WWW.REKHTA.ORG.

RESEARCH QUESTION

Given a prose excerpt from Urdu literature, find out which historical era it belongs to.

DIFFERENCE FROM PREVIOUS RESEARCH?

No one has done this before in Urdu language!

EXISTING DATASETS?

Existing datasets for Urdu prose classification lack the diachronic diversity required for this study [10, 11]

DATASET

Dividing the time period from 1800-2024 into 5 eras

STEP 1

Finding prominent literature (nasr) from each era from Rekhta and Internet Archive

STEP 2

Extracting Urdu text from PDFs using Google Vision Pro OCR (regexes and string processing), data cleaning

STEP 3

[1] Rekhta Foundation. Rekhta, www.rekhta.org.
[9] Internet Archive. Internet Archive, www.archive.org

DATASET

- **2000-2024**
  - *Ghulam Bagh* by Mirza Athar Baig
  - *Namal* by Nimra Ahmed
  - *Thanda Gosht* (Revised Editions) by Saadat Hasan Manto
- **1950-2000**
  - *Aag Ka Darya* by Qurratulain Hyder
  - *Udaas Naslain* by Abdullah Hussain
  - *Chaklawa* by Shaukat Siddiqui
- **1900-1950**
  - *Godaan* by Munshi Premchand
  - *Angaray* by Rashid Jahan and others
  - *Fasana-e-Azad* by Ratan Nath Dhar
- **1850-1900**
  - *Umrao Jan Ada* by Mirza Hadi Ruswa
  - *Fasana-e-Azad* by Ratan Nath Sarshar
  - *Mirat-ul-Uroos* by Nazir Ahmad
- **1800-1850**
  - *Bagh-o-Bahar* by Mir Amman
  - *Dastan-e-Amir Hamza* (Various versions)
  - *Aab-e-Hayaat* by Umera Ahmed

FINAL DATASET

- 15 BOOKS (3 PER ERA)
- CONVERTED FROM BOOK-WISE TO PAGE-WISE.
- EACH PAGE LABELED WITH ITS ERA
- SHUFFLING
- 80-10-10 SPLIT
- ORGANIZED INTO FOLDERS: TRAIN, VAL, TEST WITH ERA SUBFOLDERS.

Era	Train	Validate	Test
1800 - 1850	681	84	84
1850 - 1900	500	65	65
1900 - 1950	920	115	115
1950 - 2000	976	122	122
2000 - 2024	2000	250	250

COMPARISION WITH PREVIOUS WORK

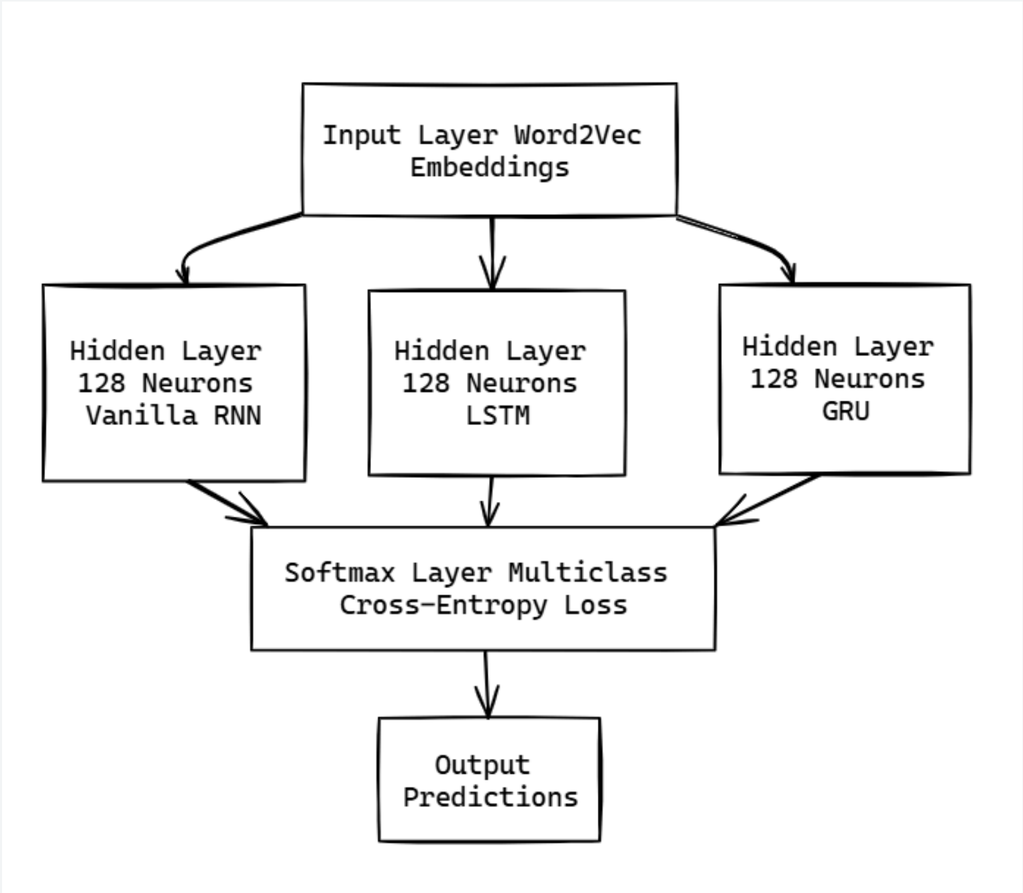
TABLE I
DIACHRONIC TEXT CLASSIFICATION RESEARCH

Year	Language	Models Used	Accuracy
2021 [1]	Hebrew	CNN, RNN (GRU), Paragraph Vectors	RNN (GRU): 85%
2019 [3]	English	Support Vector Machine (SVM)	46.3% (6-year range), 73.3% (50-year range)
2021 [4]	English	RNN	79.9%

TABLE II
URDU TEXT CLASSIFICATION RESEARCH

Year	Models Used	Accuracy
2020 [6]	Naïve Bayes, Support Vector Machines (SVM)	SVM: 93.34%, Naïve Bayes: 76.79%
2021 [7]	SVM, Decision Tree (J48), K-Nearest Neighbors (KNN)	SVM: 68.73%, Decision Tree: 62.37%, KNN: 55.41%
2022 [8]	MuRIL (BERT)	71.6%

RNN STRUCTURE



COMPARISION WITH PREVIOUS WORK

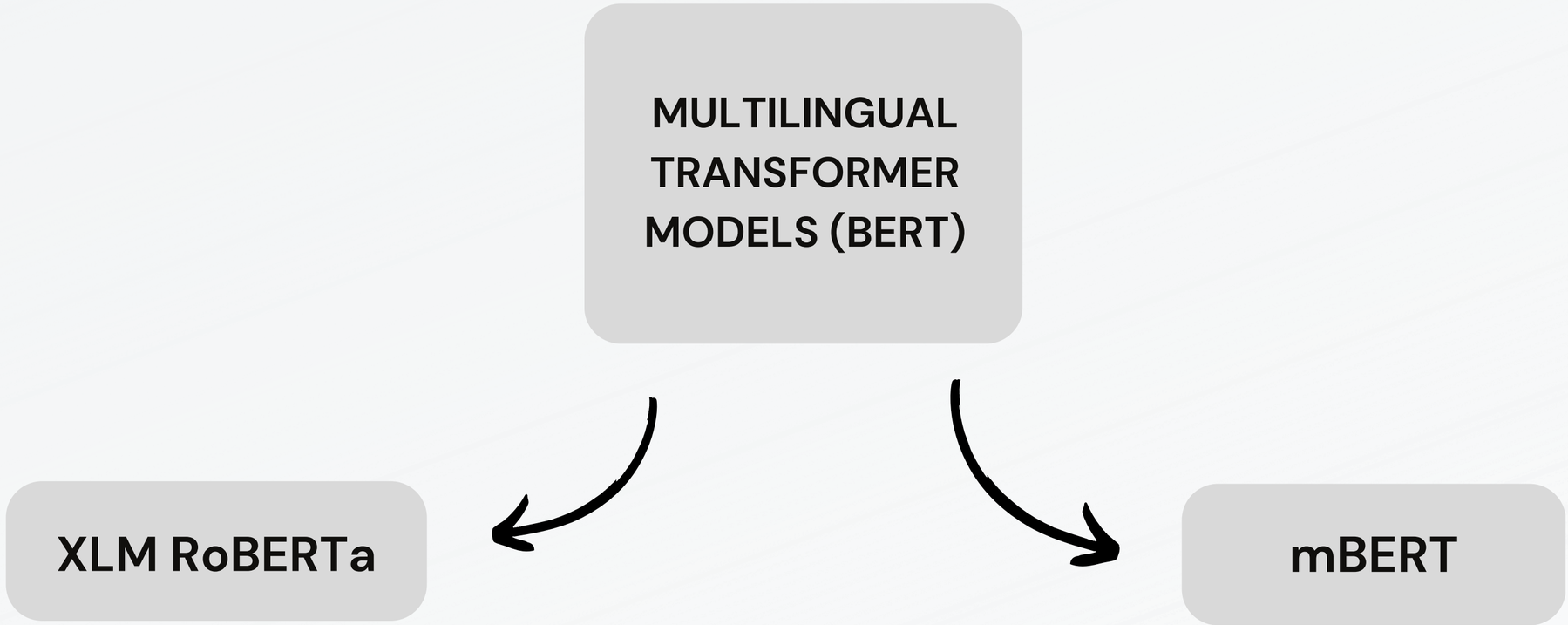
TABLE I
DIACHRONIC TEXT CLASSIFICATION RESEARCH

Year	Language	Models Used	Accuracy
2021 [1]	Hebrew	CNN, RNN (GRU), Paragraph Vectors	RNN (GRU): 85%
2019 [3]	English	Support Vector Machine (SVM)	46.3% (6-year range), 73.3% (50-year range)
2021 [4]	English	RNN	80%

TABLE II
URDU TEXT CLASSIFICATION RESEARCH

Year	Models Used	Accuracy
2020 [6]	Naïve Bayes, Support Vector Machines (SVM)	SVM: 93.34%, Naïve Bayes: 76.79%
2021 [7]	SVM, Decision Tree (J48), K-Nearest Neighbors (KNN)	SVM: 68.73%, Decision Tree: 62.37%, KNN: 55.41%
2022 [8]	MuRIL (BERT)	71.6%

FURTHER MODELS USED

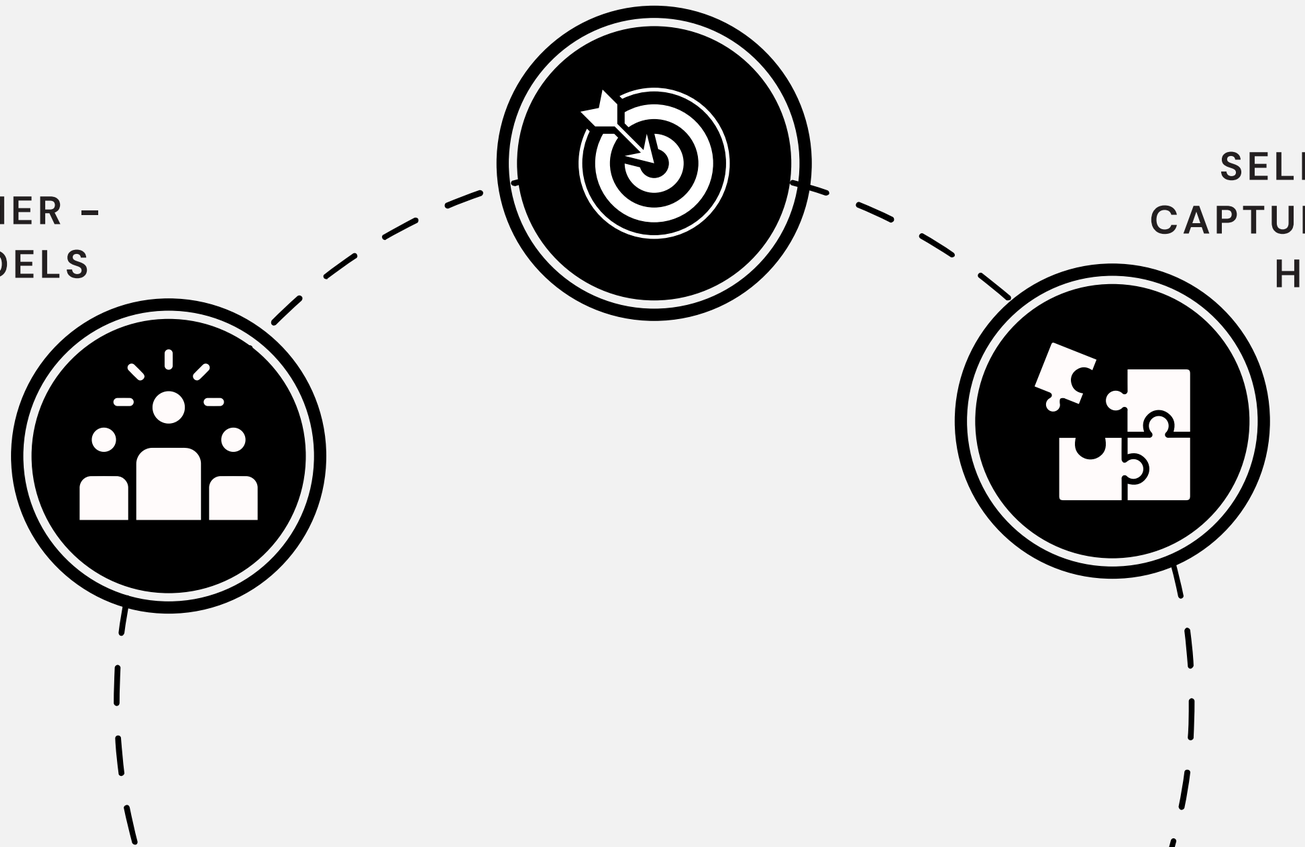


WHY THESE MODELS?

PRE-TRAINED ON MULTILINGUAL DATA

TRANSFORMER - BASED MODELS

SELF-ATTENTION HELPS CAPTURE THE DIFFERENCES IN HISTORICAL PROSE



XLM-ROBERTA

Trained on 100+ languages, including urdu

Known for superior performance in multilingual tasks

MBERT

A smaller model but equally capable

To see if a lightweight model can achieve similar results

HOW THESE MODELS WORK

Breaks prose into subwords, handling rare and historical words effectively

INPUT TOKENIZATION

Uses self-attention to capture word relationships and historical patterns

TRANSFORMER LAYERS

Final output assigns probabilities to each class (era)

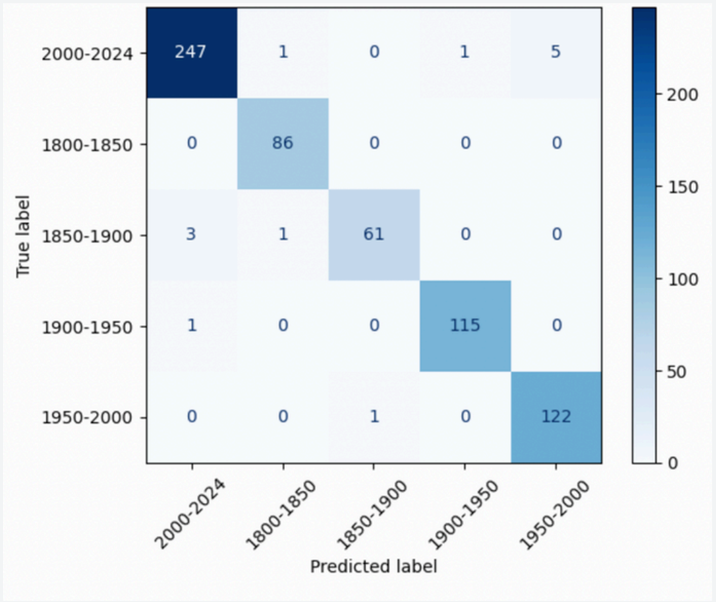
OUTPUT CLASSIFICATION

HYPERPARAMETERS

EPOCHS	3
BATCH SIZE	16
LEARNING RATE	0.00002
LOSS FUNCTION	DEFAULT (XLM R)
LABELS	5 (ERAS)
MAX SEQUENCE LEN	512

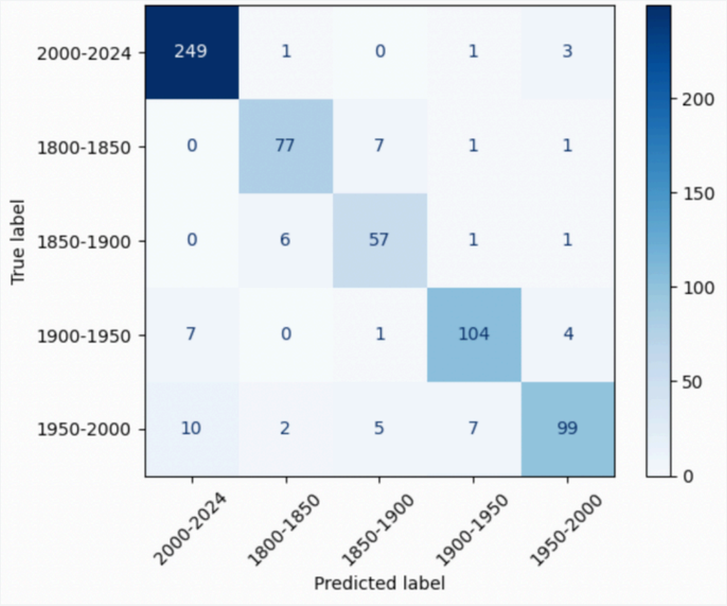
RESULTS

XLM-ROBERTA



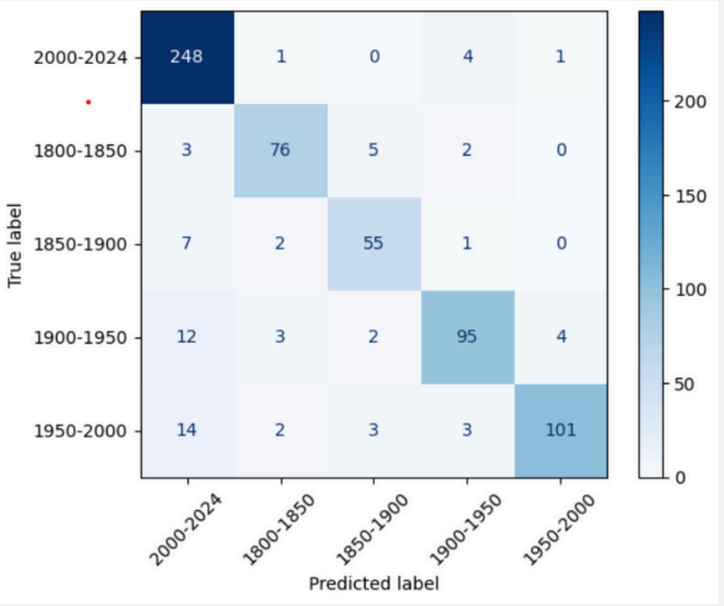
Accuracy: 98%

M-BERT



Accuracy: 91%

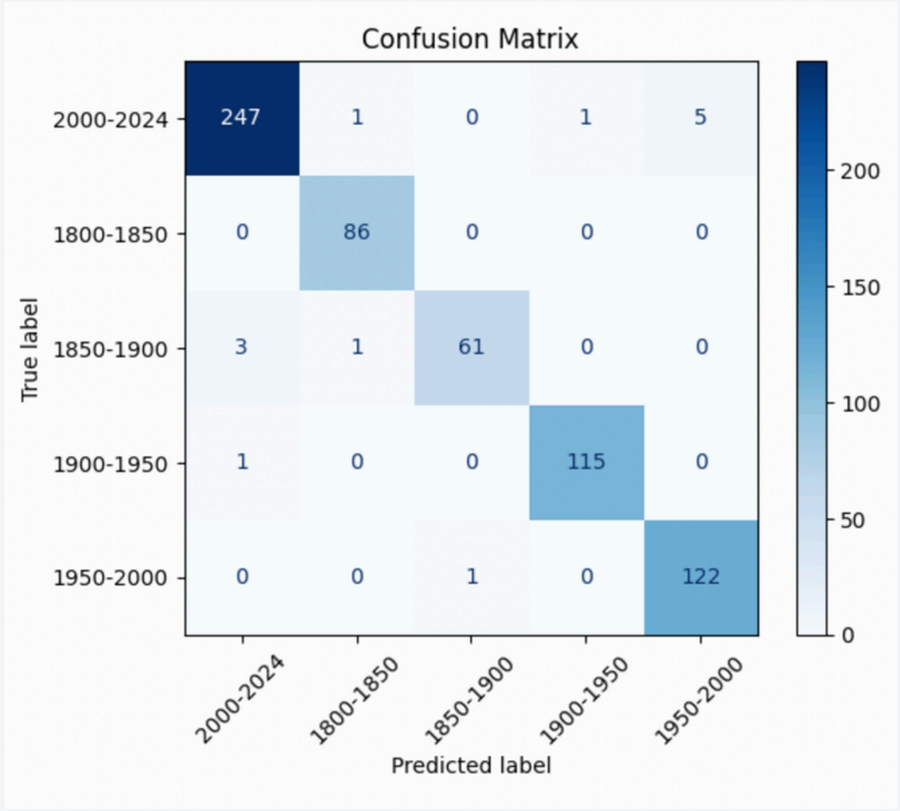
RNN



Accuracy: 79%

RESULTS

XLM-ROBERTA



Accuracy: 98%.

WHY XLM-ROBERTA PERFORMED BETTER:

- LARGER PRE-TRAINING DATASET.
- EXCELLENT AT UNDERSTANDING COMPLEX RELATIONSHIPS IN TEXT, BENEFICIAL FOR HISTORICAL PROSE.

RNN:

- LACK THE PRETRAINING ADVANTAGE AND GLOBAL ATTENTION MECHANISM OF MODELS LIKE XLM-ROBERTA AND MBERT

ADVANCES OVER PREVIOUS RESEARCH

- FIRST TO EVER DO IT IN URDU
- USED BERT FOR TEMPORAL CLASSIFICATION (NO TRANSFORMERS USED BEFORE)
- **RNN:** ENGLISH: 79.9%, HEBREW: 85%
- BETTER ACCURACY FOR RNN USING SAME STRUCTURE

ACHIEVEMENTS AND FUTURE WORK

ACHIEVEMENTS

- ACHIEVED GOOD ACCURACY IN A NEW DOMAIN
- GENERATED DATASET
- PROSE WITHOUT DATES CAN BE CLASSIFIED

FUTURE WORK

- A LARGER DATASET
- TRY MORE MODELS (SVMS)

REFERENCES

- [1] REKHTA FOUNDATION. REKHTA, WWW.REKHTA.ORG. ACCESSED 15 OCT. 2024.
- [2] LIEBESKIND, CHAYA, AND SHMUEL LIEBESKIND. "DEEP LEARNING FOR PERIOD CLASSIFICATION OF HISTORICAL TEXTS." JOURNAL FOR DATA MINING AND DIGITAL HUMANITIES, 2019. HAL ID: HAL-02324617, VERSION 1 SUBMITTED ON 22 OCT. 2019, LAST REVISED 1 JUNE 2020 (VERSION 2), HTTPS://HAL.SCIENCE/HAL-02324617V1.
- [3] HE, YU, ET AL. "TIME-EVOLVING TEXT CLASSIFICATION WITH DEEP NEURAL NETWORKS." PROCEEDINGS OF THE TWENTY-SEVENTH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI-18), IJCAI, 2018, PP. 2014-2020. INTERNATIONAL JOINT CONFERENCES ON ARTIFICIAL INTELLIGENCE ORGANIZATION, STOCKHOLM, SWEDEN, WWW.IJCAI.ORG/PROCEEDINGS/2018/0279.PDF.
- [4] JO, EUN SEO, AND MARK ALGEE-HEWITT. "THE LONG ARC OF HISTORY: NEURAL NETWORK APPROACHES TO DIACHRONIC LINGUISTIC CHANGE." JOURNAL OF THE JAPANESE ASSOCIATION FOR DIGITAL HUMANITIES, VOL. 3, NO. 1, 2021, P. 32. WWW.JSTAGE.JST.GO.JP/ARTICLE/JJADH/3/1/3_1/_PDF. ACCESSED 15 OCT. 2024.
- [5] SZYMANSKI, TERRENCE, AND GERARD LYNCH. "UCD: DIACHRONIC TEXT CLASSIFICATION WITH CHARACTER, WORD, AND SYNTACTIC N-GRAMS." PROCEEDINGS OF THE 9TH INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION (SEMEVAL 2015), ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, JUNE 2015, PP. 879-883, DENVER, CO, USA. DOI:10.18653/V1/S15-2148, WWW.SEMANTICSCHOLAR.ORG/PAPER/UCD-%3A-DIACHRONIC-TEXT-CLASSIFICATION-WITH-WORD%2C-AND-SZYMANSKI-LYNCH/9A286BA134489E0C4173BA2522A1D71CEFABB209.

REFERENCES

- [6] ALI, ABBAS, AND MALIHA IJAZ. "URDU TEXT CLASSIFICATION." PROCEEDINGS OF THE ACM INTERNATIONAL CONFERENCE, DEC. 2009, P. 21. DOI:10.1145/1838002.1838025.
- [7] RASHEED, IMRAN, ET AL. "URDU TEXT CLASSIFICATION: A COMPARATIVE STUDY USING MACHINE LEARNING TECHNIQUES." 2018 THIRTEENTH INTERNATIONAL CONFERENCE ON DIGITAL INFORMATION MANAGEMENT (ICDIM), 2018, PP. 274–278. DOI:10.1109/ICDIM.2018.8847044.
- [8] BHAUMIK, A. B., & DAS, M. (2022). “EMOTIONS & THREAT DETECTION IN URDU USING TRANSFORMER-BASED MODELS”. IN FIRE’22: FORUM FOR INFORMATION RETRIEVAL EVALUATION, DECEMBER 9–13, 2022, INDIA (PP. 1–8). CEUR WORKSHOP PROCEEDINGS. HTTP://CEUR-WS.ORG/
- [9] INTERNET ARCHIVE. INTERNET ARCHIVE, WWW.ARCHIVE.ORG. ACCESSED 15 OCT. 2024.
- [10] HUANG, XIAOLEI, AND MICHAEL J. PAUL. "NEURAL TEMPORALITY ADAPTATION FOR DOCUMENT CLASSIFICATION: DIACHRONIC WORD EMBEDDINGS AND DOMAIN ADAPTATION MODELS." PROCEEDINGS OF THE 57TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, JULY 2019, PP. 4113–4123, FLORENCE, ITALY. WWW.ACLANTHOLOGY.ORG/P19-1403.PDF.
- [11] KHAN, WAHAB, ET AL. "NAMED ENTITY DATASET FOR URDU NAMED ENTITY RECOGNITION TASK." JOURNAL OF INFORMATION SCIENCE AND ENGINEERING, JAN. 2016.

THANK YOU

Open for Questions and Discussion