Predicting Age from Social Media Language

Group 5: Angel Vasquez, Eliana Durell, Max Frohman

Project Overview

Research Question: What linguistic features can be leveraged to predict a writer's age or age group in an age prediction NLP task?

- Motivation:
  - Moderation and Safety
  - Marketing and Audience Targeting
  - Security and Threat Detection
  - Research and Data Labeling

Hypothesis: Using a transformer-based approach will pick up on both broader structural features (syntax) and more fine-grained age-indicating language usage like specific topics, punctuation trends, and slang

Keywords: age classification, text classification, social media, blogging

Age and gender in language, emoji, and emoticon usage in instant messages

Koch et al., 2022

Key Points

Older people

- Theme
- Less emojis
- Longer messages
- Proper punctuation
- Greetings

Younger people

- More emojis
- Informal expressions
- First person

Methodology

- WhatsApp messages from volunteers
- Used an open and closed vocabulary approach
- Tokenization
  - Filtering
- Models
  - Elastic Net
  - Random Forest
  - Baseline

Relevance

- Features they used
- Open vocabulary approaches were superior (data driven)
- Model standings
- Topics

TextAge: A Curated and Diverse Text Dataset for Age Classification

Cheekati et al., 2024

Key Points

- Created a complete dataset from relevant sources
- High accuracy in detecting underage vs adults
- Did well at identifying younger groups but struggled with older generations

Methodology

- Utilized RoBERTa and XLNet models
- Fine tuned pre trained models to better understand
-

Relevance

- Shows the usefulness of pretrained models like RoBERTa
- Cautions us about age representation in data
-

Predicting age groups of Twitter users based on language and metadata features
Morgan-Lopez et al., 2017

Key Points

- Identifies a need for non-proprietary demographics classification for any group
- Finds it is easier to distinguish between ages 13-17 and 18-24
- Linguistic and metadata features show only a slight improvement to just linguistic features

Methodology

- Classical machine learning techniques on pre-extracted features
- Created models for language features, metadata features, both, and WWBP words
- 10-fold cross-validation and multiple forms of loss

Relevance

- Demonstrates a clear need for more accurate and sophisticated age classification
- Strong performance for language-based features on our target age groups (minors and young adults)

Demographic Inference on Twitter using Recursive Neural Networks

Kim et al., 2017

Key Points

- Standard supervised learning approaches on text features do not take advantage of the depth of the data
- Graph RNNs outperform other ML techniques for age and gender classification

Methodology

- Uses Graph RNNs
  - RNNs provide better language processing
  - Graph/tree structure accounts for network topology present in Twitter user data
- Compare with five other ML techniques

Relevance

- Continuing to deepen the models used for demographic classification leads to better performance
- Graphs are meant to capture metadata we do not want at the cost of high memory usage – we can go deeper, not wider

Dataset: Blog Authorship Corpus

- From “Effects on Age and Gender on Blogging” by Schler et al., 2006
- Collection of over 600,000 posts from over 71,000 blogs on blogger.com as of 2004
- Age labels for posts spanning between 13 and 48

	gender		
age	**female**	**male**	**Total**
unknown	12287	12259	24546
13-17	6949	4120	11069
18-22	7393	7690	15083
23-27	4043	6062	10105
28-32	1686	3057	4743
33-37	860	1827	2687
38-42	374	819	1193
43-48	263	584	847
>48	314	906	1220
Total	34169	37324	71493

Table 1 Blogs Distribution over Age and Gender

Classed as →	**10's**	**20's**	**30's**
10's	7036	1027	177
20's	916	6326	844
30's	178	1465	1351

Table 7 Confusion matrix for the age classifier using all features

Methodology and Metrics

Model

- RoBERTa
- LIME

Tools

- PyTorch
- Hugging Face Transformers
- spaCy

Evaluation

- Baseline models – pretrained RoBERTa
- LIME for ablation study
- Custom train, validation, and test splits on our data

Metrics

- Sklearn for metrics and visualization
- F1-score, confusion matrix
- Compare with baseline models and original dataset study

Risks

Our data covers a fairly large age range but is skewed towards late-teens/early-twenties and only contains users aged 13-48, so our final model may contain age biases.
Furthermore, because our dataset is older (around 20 years old), it does not accurately capture current slang and language trends, which could result in mistaken classifications or poor accuracy on modern data.

Contributions

Prior research shows that language models trained on Facebook posts are more accurate because people self-disclose more (Koch et al. 2022). Because our dataset has lots of self-disclosed data, it could result in a model that is very accurate for age classification. Better age classification models will improve social media moderation.

Questions?