# LP 2

50% of paper can be joined work
it has to be identifiable who wrote what

# Project task

Patrizia Paggio

paggio@hum.ku.dk

# Chosen task for project

## From the CLEF 2024 challenges:

- Multi-Author Writing Style Analysis 2024

https://pan.webis.de/clef24/pan24-web/style-change-detection.html

# Synopsis

- Task: Given a document, determine at which positions the author changes.
- Input: Reddit comments, combined into documents.
- Output: Where does authorship change on the paragraph level?
- Evaluation: F1 (code provided).

# Task

**For a given text, find all positions of writing style change on the paragraph-level** (i.e., for each pair of consecutive paragraphs, assess whether there was a style change).

All documents are in English and may contain an arbitrary number of style changes.

Style changes may only occur between paragraphs.

# Task

**Three difficulty levels**:

1. **Easy:** Mix of topics in same doc, allowing use of topic information to detect authorship changes.
2. **Medium:** Small topical variety in a document. Approaches to focus more on style to solve the detection task.
3. **Hard:** All paragraphs in a document are on the same topic.

Exam!!!

# Task

## Think break

Focus is on the difference between topic and style.

How would you define this difference?

Topic: Words that are used (not register but cluster of words, words that cluster thematically, choose a name for topic based on cluster centroid?)

Style: syntax, punctuation (did they remove punctuation in preprocessing??), register, specific words that occur more frequently, sentence length, paragraph length?

# Data

Three datasets, one for each level of difficulty.

Each dataset is split into three parts:

training set: Contains 70% of the whole dataset and includes ground truth data.

validation set: Contains 15% of the whole dataset and includes ground truth data.

(test set: Contains 15% of the whole dataset, no ground truth data is given).

# Data

Additional external data for training can be used.

However, the additional data used must be made freely available under a suitable license.

# Data

*** Datasets for the exam ***

You will use the training/development data (The challenge test data are not yet available and **will not be included**).

You will decide on an evaluation strategy (k-fold cross validation?).

# Data

## Input format

For each problem instance X (i.e. each input document), two files are provided:

1. problem-X.txt contains the actual text.
2. truth-problem-X.json contains the ground truth, i.e., the correct solution in JSON format.

# Data

## Ground truth

Example of a multi-author document with a style change between the third and fourth paragraph:

{
"changes": [0,0,1,...]
}

# From last year's task

The task was the same as this year, and there were three different levels of difficulty.

First time Reddit data were used. Subreddits were selected: *r/worldnews*, *r/politics*, *r/askhistorians*, and *r/legaladvice*.

All documents were created by 2-4 authors.

Total of 6000 documents.

# From last year's task

## Intrinsic and extrinsic models

Intrinsic: based on paragraph/sentence analysis

Extrinsic: finds the authors' names online (to show a weakness in the methodology)

???

# From last year's task

## Submissions

6 groups submitted, all adopting intrinsic methods.

**Table 1**
Overall results for the style change detection task. The best result for each data set is given in bold.

| Participant | Easy $F_1$ | Medium $F_1$ | Hard $F_1$ |
|---|---|---|---|
| Chen et al. [35] | 0.914 | 0.820 | 0.676 |
| Hashemi et al. [41] | **0.984** | **0.843** | 0.812 |
| Huang et al. [38] | 0.968 | 0.806 | 0.769 |
| Jacobo et al. [40] | 0.793 | 0.591 | 0.498 |
| Kucukkaya et al. [37] | 0.982 | 0.810 | 0.772 |
| Ye et al. [33] | 0.983 | 0.830 | **0.821** |

# From last year's task

Hashemi et al. apply the BERT, the RoBERTa and the ELECTRA pre-trained language models which they combine with a binary classification layer.

Similar to Chen et al., they also apply data augmentation strategies to generate additional training data, and experiment with ensembles of models trained on all three data sets.

Ye at al. use the pre-trained large language model DeBERTaV3 combined with prompting, essentially asking the language model whether two paragraphs are written by the same author or not.

# Style modelling

- Lexical character/word features
- Stylometric features
- Syntactic/semantic features
- Structural features
- Language models

# Style modelling

The usefulness of different stylometric features depend on text length, text genre, or **topic variation**. (Stein et al. 2010)

# Lexical features

- Counts of adjacent words (unigrams, bigrams, trigrams, …)
- or adjacent characters

What do n–grams model?

Vocabulary more than style (especially word n-grams).

# Lexical features

## Character n-grams

Advantages?

- Language independent.
- Tolerant to noise.

"simplistic" and "simpilstc" will generate many similar trigrams.

# Lexical features

They can be used to be build a bag-of-word model:

vectors where each n-gram is represented as either 0 or 1.

# Lexical features

They can be combined with weighting, for instance as given by the TF-IDF weighting scheme:

vectors where each n-gram is represented by a score.

TF($t,d$) = relative frequency of term $t$ in doc $d$

IDF($t,N$) = log($N/n$)

where $N$=#docs, and $n$=#docs in which $t$ appears

# Stylistic/stylometric features

Any suggestions?

(remember the first practical we did for this course)

# Stylistic/stylometric features

- Avg. sentence and word length
- Type/token ratio
- Repeated errors (typos)
- Hashtag counting
- Use of punctuation
- …

(remember the first practical we did for this course)

# Stylistic/stylometric features

- Avg. sentence and word length

These can be applied to any language.

We just need a tokeniser.

# Stylistic/stylometric features

- Type/token ratio V/N (**V**ocabulary size, **N**umber of tokens)

But **careful**:

V depends heavily on text length, and words tend to get repeated in longer texts.

The relationship between V and N is not linear.

Okay if texts being compared are roughly equal in length.

# Stylistic/stylometric features

- Repeated writing errors (typos)

Capture the idiosyncrasies of an author's style.

”Man is the only animal that trips twice over the same stone”.

# Stylistic/stylometric features

Text bleaching (van der Goot et al. 2018: 384)
to remove content from the representation.

| Original | a | bag | of | Doritos | for | lunch! | ¯_(ツ)_/¯ |
|---|---|---|---|---|---|---|---|
| Frequency | 4 | 2 | 4 | 0 | 4 | 1 | 0 |
| Length | 01 | 03 | 02 | 07 | 03 | 06 | 04 |
| PunctC | W | W | W | W | W | W! | ¯_(ツ)_/¯ |
| PunctA | W | W | W | W | W | WP | JJJJ |
| Shape | L | LL | LL | ULL | LL | LLX | XX |
| Vowels | V | CVC | VC | CVCVCVC | CVC | CVCCCO | OOOO |

Table 1: Abstract features example transformation.

# Syntactic and semantic features

- POS-tags (and n-grams building on them)

They add more specificity to word-based n-grams.

Quality depends on resources for the specific language.

# Syntactic and semantic features

- Constituent structure

It can be used as a measure of complexity with measures such as avg. branching factor, avg. height for noun or verb phrases.

We found a use for parse trees!

# Syntactic and semantic features

- Dependency structure

It can be used to construct vectors of labelled relations that represent the structured content of the sentences.

Question: how well does dependency parsing work on Reddit comments?

# Language models

Several possibilities exist:

- Word2vec
- GloVe
- FastText
- BERT
- …

# Language models

Different embedding types have different advantages

| Embedding | Advantage |
|---|---|
| Word2vec | easy to use, fast |
| GloVe | better word co-occurrence stats |
| FastText | takes into account the internal structure of words, better for unknown/mispelt words |
| BERT | accounts for different senses of a word |

# What happens now?

You will experiment with the specific dataset and discuss the use of various features&classifiers in a 'method brainstorming workshop', which is planned for April 26.

For the exam we will ask you to work at least with the middle level.