**Problem Formulation Paper**

Topic:

**Classification of the Dataset: One Million Post Corpus**

Course:

Natural Language Processing

Department of Mathematics and Computer Science

At Phillips Marburg University

Submitted by Team10:

*Tim Luckhardt, Mtr. 3142907*

*Samuel Becker, Mtr. 2847416*

*Mehmed Yilmaz, Mtr. 3476521*

Lecturers:

Prof. Dr. Lucie Flek

Vahid Sadiri Javadi

Paper submission date

29.05.2022

1. Introduction

For the following Group Project in the Course Natural Language Processing we decided to train a classification Model on the Dataset One Million Post Corpus.

This Dataset contains annotated user posts from the Austrian Newspaper Website “Der Standard”.

The Main objective for our project is to train a model that is able to predict the annotated categories mentioned in Chapter 1.2 .

As subtask we would like to train another model that is able to predict the label status which helps to identify content that should be deleted and therefore probably has a relation to the negative associated categories.

Eventually, given the time constraints and project scope enables us, we would like to take a closer look at the distance between categories to find out how they correlate and how the categories correlate to the topic path of news articles.

Table Of Content

[I. Introduction 2](#_Toc104325578)

[1 Dataset Description 4](#_Toc104325579)

[1.1 Dataset Labels 4](#_Toc104325580)

[1.1.1 Post Labels 4](#_Toc104325581)

[1.1.2 Article Labels 4](#_Toc104325582)

[1.2 Dataset Annotations for user generated content 5](#_Toc104325583)

[1.2.1 Potentially undesirable content 5](#_Toc104325584)

[1.2.2 Neutral content that requires a reaction 5](#_Toc104325585)

[1.2.3 Potentially desirable content 5](#_Toc104325586)

[1.3 Dataset Issues 5](#_Toc104325587)

[2 The Do-ing 6](#_Toc104325588)

[2.1 Prediction of post categories 6](#_Toc104325589)

[2.2 Prediction of label status 6](#_Toc104325590)

[3 Summary 6](#_Toc104325591)

# Dataset Description

The “One Million Posts” corpus is an annotated data set consisting of user comments posted to the Austrian newspaper website “Der Standard” (in German language).

Der Standard is an Austrian daily broadsheet newspaper. On the newspaper’s website, there is a discussion section below each news article where readers engage in online discussions. The data set contains a selection of user posts from the 12 month time span from 2015-06-01 to 2016-05-31. There are 11,773 labeled and 1,000,000 unlabeled posts in the data set. The labeled posts were annotated by professional forum moderators employed by the newspaper.

## Dataset Labels

### Post Labels

The data set contains the following data for each post:

* Post ID

→ 206

* Article ID

→ 4

* Headline (max. 250 characters)

→ “Ja aber: Was hat der Stance….”

* Main Body (max. 750 characters)

→ “Ich bin Rechtshänder, fahre Goofy und kicke mit Rechts. Sie sind Linkshänder, ..”

* User ID (the user names used by the website have been re-mapped to new numeric IDs)

→ 10467

* Time stamp

→ “2014-08-13 08:00:52.207”

* Parent post (replies give rise to tree-like discussion thread structures)

→ 203

* Status (online or deleted by a moderator)

→ “online”

* Number of positive votes by other community members

→ 0

* Number of negative votes by other community members

→ 0

### Article Labels

For each article, the data set contains the following data:

* Article ID

→ 4

* Publishing date

→ “2014-08-13 05:30:00.00”

* Topic Path (e.g.: Newsroom / Sports / Motorsports / Formula 1)

→ “Newsroom/User/mitmachen/Mitreden”

* Title

→ “Welche Erfahrungen haben Sie als Linkshänder gemacht?”

* Body

→ “<div class="section" … Wie sieht Ihr Alltag als Linkshänder aus? Erledigen Sie manche Arbeiten mit rechts statt mit links? ...”

## Dataset Annotations for user generated content

### Potentially undesirable content

* Sentiment (negative/neutral/positive)

An important goal is to detect changes in the prevalent sentiment in a discussion, e.g., the location within the fora and the point in time where a turn from positive/neutral sentiment to negative sentiment takes place.

* Off-Topic (yes/no)

Posts which digress too far from the topic of the corresponding article.

* Inappropriate (yes/no)

Swearwords, suggestive and obscene language, insults, threats etc.

* Discriminating (yes/no)

Racist, sexist, misogynistic, homophobic, antisemitic and other misanthropic content.

### Neutral content that requires a reaction

* Feedback (yes/no)

Sometimes users ask questions or give feedback to the author of the article or the newspaper in general, which may require a reply/reaction.

### Potentially desirable content

* Personal Stories (yes/no)

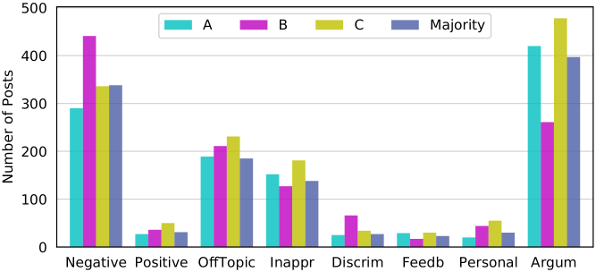
In certain fora, users are encouraged to share their personal stories, experiences, anecdotes etc. regarding the respective topic.

* Arguments Used (yes/no)

It is desirable for users to back their statements with rational argumentation, reasoning and sources.

## Dataset Issues

Imbalanced data between catgeories.



# The Do-ing

## Prediction of post categories

Training a model that can predict the categories of posts mentioned in Chapter 1.2.

For this task we will load the Dataset into a classifier and split into training and test data.

We will not use accuracy for performance measurement because of data imbalance. Instead, we will be using Mean Squared Error and F1 Score to compare the performance.

MSE will give us an idea of how far our prediction is off and F1 Score will give us an idea of what prediction is wrong.

## Prediction of label status

The label status has only two values. Therefore, we can use binary classification solutions.

For measuring the performance, we will be using the metric Area Under the Curve (AUC).

AUC has a range of [0, 1]. The greater the value, the better is the performance of our model.

# Summary

We are given a dataset from a newspaper with editorially and user generated content, on which we would like to train a model to predict the category of user generated content as well as status of that content.

Performance measures are MSE & F1-Score for task one and AUC for task two.