University of Regensburg

Institute for Language, Literature and Cultural Studies

Chair of Media Informatics

****

Exposé for a bachelor thesis on the topic:

**CookBERT – Cooking with BERT**

Domain-adaptive pretraining of BERT on cooking conversation data

**Submitted by:**

Pascal Strobel

Gluckstraße 3

93053 Regensburg

E-Mail: pascal.strobel@stud.uni-regensburg.de

Matriculation number: 2106133

Bachelor thesis in Media Informatics

7. Fachsemester

Supervisor: Mr. Alexander Frummet (M. Sc.)

First reviewer: Prof. Dr. Udo Kruschwitz

Second reviewer: PD Dr. David Elsweiler

Datum: 24.11.2021

Inhaltsverzeichnis

[Problem definition 2](#_Toc88745093)

[Objective and interest 3](#_Toc88745094)

[Forschungsstand und theoretische Grundlagen 4](#_Toc88745095)

[Forschungskonzept 5](#_Toc88745096)

[Vorläufige Gliederung 6](#_Toc88745097)

[Rough time schedule 7](#_Toc88745098)

[Literaturverzeichnis 8](#_Toc88745099)

# Problem definition

The introduction of large-scale pre-trained transformers has led to a small revolution in the field of NLP. Such language models (LMs) are pre-trained on a massive amount of unlabeled text data with self-supervised objectives (e. g. MLM, next sentence prediction, ...) and achieve state-of-the-art performance on a wide variety of tasks, including text classification, named entity recognition, question answering and machine translation. However, the pre-training data mostly originates from general topics and domains (e.g. Wikipedia articles and books corpus for BERT), leading to limited knowledge of the LM about domain-specific vocabulary or text structures (like dialogues).

To counteract this lack of domain knowledge, different approaches can be found in the literature, which basically all try to do so by feeding domain- or task-specific data into the model (be it by pre-training a model from scratch, or by further pre-training an already existing model). Not surprisingly, such approaches generally outperform the standard pre-trained models when applied to tasks in their domain.

While meanwhile an extensive number of domain-specific language models exists, there is not yet a model that is geared towards cooking conversation data and thus is a candidate for a conversational agent for the kitchen.

# Objective and procedure

The main goal of this work ist to answer the following research question:

„How does domain-adaptive pre-training of BERT on cooking conversation data affect the performance of downstream tasks relevant to conversational agents in the kitchen?“

To answer this question, an already pre-trained LM (BERT) is enriched with domain-specific knowledge about cooking conversations. This is done via domain-adaptive pre-training, which is nothing but further pre-training the model on domani-specific data on the two target objectives: masked language modeling and/ or next sentence prediction. Since no matching data is publicly available, a textual data set of unlabeled cooking conversations has first to be created.

The model is then finetuned and evaluated on (several) downstream tasks that are relevent for conversational agents in the kitchen, e. g. intent classification and named entity recognition.

* It is still unclear, how many (and what) tasks the model will be finetuned and tested on.

By a comparison with other, already existing models (that are directly finetuned and thus do not have the additional domain knowledge), the research question can then be answered.

* Comment: Pellegrini et al. (2021) trained on the Recipe1M dataset via domain-adaptive pre-training (i.e., on food data in general, and not on cooking conversation data as I will do). However, since this domain is quite similar to my target domain, a performance comparison of the two models would also be very interesting!

**Summary of the objectives of the work:**

* Creation of an unlabeled, textual cooking conversation dataset
* Domain-adaptive pre-training of an existing LM (BERT)
* Finetuning and evaluating the model on relevant downstream tasks
* Model comparison and answering research question

It is expected that the proposed model with domain-specific knowledge will outperform existing, general models on all conversational agent relevant tasks.

# Forschungsstand und theoretische Grundlagen

* adf
* TAPT und DAPT in der Literatur inzwischen weit verbreitet und liefert sehr zuverlässige Ergebnisse

In sozialen Gruppen existieren Meinungsführende (Opinion-Leaders), die als Fachleute sowie Ratgebende für bestimmte Themen gelten und die im Rahmen interpersonaler Kommunikation ihre Bewertungen, Interpretationen oder Meinungen an die Meinungsfolgenden (Opinion-Followers) weitergeben (Lazarsfeld, Berelson & Gaudet, 1948). Daneben bezeichnen ‚virtuelle Meinungsführer‘ Medienpersönlichkeiten, die hohes Vertrauen und Ansehen genießen, mit denen jedoch nicht interpersonal kommuniziert wird (Katz, 1957).

Infolge der erhöhten Nutzung sozialer Medien wird die Gültigkeit der Theorie jedoch infrage gestellt: Bennett und Manheim (2006) stellten fest, dass die Grenzen zwischen den Medien, den Meinungsführenden und der Öffentlichkeit bei sozialen Medien verschwimmen. Thorson und Wells (2015) entwarfen daher ein neues Modell, das explizit den Informationsfluss in sozialen Netzwerken beschreibt.

Wie sich der Two-Step-Flow und das Konzept des Opinion-Leadership auf Twitter anwenden lassen, wurde noch nicht in der Literatur behandelt und der Aspekt der politischen Kommunikation wurde ebenfalls weitgehend vernachlässigt.

# Forschungskonzept

Die folgenden Fragen sollen in dieser Arbeit beantwortet werden:

* Entspricht Trumps Tweetverhalten dem eines klassischen Opinion-Leaders?
* Wer gehört zu Trumps Opinion-Followers?
* Wie gestaltet sich der Two-Step-Flow, wenn Trump bspw. negativ über Unternehmen tweetet?

Um den Two-Step-Flow zu visualisieren, wird der Fall Nordstrom betrachtet. Hierbei handelt es sich um eine Kaufhauskette, die sich gegen die Weiterführung der Modelinie von Trumps Tochter Ivanka entschied. Kurze Zeit später veröffentlichte Donald Trump einen Tweet darüber, dass das Unternehmen seine Tochter ungerecht behandelt habe.

Im Rahmen der Bachelorarbeit werden die folgenden methodischen Schritte umgesetzt:

1. Eine Case-Study wird durchgeführt, um zu untersuchen, welche Auswirkungen der Tweet auf die Medienlandschaft hatte.
2. Es wird analysiert, wie das Weiße Haus daraufhin kommunizierte und wie die Nachricht von den Medien aufgegriffen wurde.
3. Der Two-Step-Flow wird visualisiert und der Agenda-Building-Prozess untersucht.

Dabei wird die Methodik der qualitativen Inhaltsanalyse verwendet, um die Kommunikation in den sozialen Medien im Fall Nordstrom zu analysieren. Als Datensatz dienen alle relevanten Beiträge (Tweets) auf der Plattform Twitter, in denen über den Fall Nordstrom kommuniziert wird.

# Preliminary outline

1. Introduction (incl. research question)
2. Objectives
3. Related Work
4. Creation of a cooking conversation dataset
   1. Origin and properties of the gathered data

* Where does the data come from (cooking tv shows, podcasts, …) and what kind of data is it?
  1. Preparation of the data
* How was data prepared for the domain-adaptive pre-training?
  1. Advantages and limitations of the data

1. Explanation of the base model used (BERT)

* Explain the general structure of BERT

1. Domain-adaptive pre-training
   1. Pre-training objectives

* Masked language modelling and next sentence prediction
  1. Data preprocessing
* Tokenizing etc.
  1. Pre-training specifications
* What optimizer was used? how many training epochs? Handling of over-/underfitting?

1. Finetuning
   1. Task 1

* Brief explanation of task 1 (e. g. intent classification) and corresponding dataset
  1. Task 2
  2. …

1. Evaluation
   1. Models for comparison

* Brief explanation, what (base) models are used for the comparison
  1. Comparison of model performances

1. Conclusion
2. Summary

# Rough time schedule

**Available time**: Sieben Wochen (01.01.2021–19.02.2021)

* Already done: Literature research
* Already done: Data collection (for domain-adaptive pre-training)
* Until 03.12.: Data preparation
* Until 10.12.: Familiarizing with huggingface/ FARM library
* Until 24.12.: Domain-adaptive pre-training of existing BERT model (via huggingface or FARM by deepset-ai)
* Until 14.01.: Finetuning on different downstream tasks
* Until 28.01.: Evaluation
* Until 18.02.: Writing (also already during the practical parts)
* Until 25.02: Test reading and finetuning
* Until 01.03.: Submission

# Literaturverzeichnis

Bennett, W. L., & Manheim, J. B. (2006). The One-Step Flow of   
Communication. *The ANNALS of the American Academy of Political and Social Science*, 608(1), pp. 213–232.

Crockett, Z. (2016, 16. Mai). *What I learned analyzing 7 months of Donald Trump’s tweets*. Vox. https://www.vox.com/2016/5/16/11603854/donald-trump-twitter

Katz, E. (1957). The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *Public Opinion Quarterly*, *21*(1), 61–78. https://doi.org/10.1086/266687

Katz, E., & Lazarsfeld, P. (1955). *Personal Influence: The Part Played by People in the Flow of Mass Communications.* New York: Free Press.

Lazarsfeld, L., Berelson, B., & Gaudet, H. (1948). *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Columbia University Press.

Ott, B. L. (2016). The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication*, *34*(1), 59–68. https://doi.org/10.1080/15295036.2016.1266686

Thorson, K. & Wells, C. (2015). Curated Flows: A Framework for Mapping Media Exposure in the Digital Age. *Communication Theory*, *26*(3), 309–328. https://doi.org/10.1111/comt.12087

References

* Pellegrini, C., Özsoy, E., Wintergerst, M., & Groh, G. (2021, February 11–13). Exploiting Food Embeddings for Ingredient Substitution. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies* (pp. 67–77). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0010202000670077