**Plan of work (appended to thesis contract)**

Name of student Hisham Alkaed

Degree programme Text Mining M.Sc.

Student number 2671370

Name of supervisor(s) Ilia Markov

Chair VU Amsterdam, Faculty of Humanities, CLTL lab, Assistant Professor.

Second reader/ Antske Fokkens

Chair VU Amsterdam, Faculty of Humanities, CLTL lab, Professor.

Working title Exploring the Role of Textual Modality in Hateful Meme Detection

Topic The hateful meme detection task presents a unique problem in the field of text mining and image recognition. This problem requires advanced algorithms and techniques to analyze and extract meaningful information from multimodal features, effectively combining image and text data to identify and classify hateful content in memes. Within this context the rule of textual modality will be studied.

Aim and relevance Studying the textual aspect of hateful memes, which have become a popular way for people to spread hate speech, has numerous benefits for improving online safety through introducing useful insights and methods to automatically detect that in online platforms. In addition, it could help raise public awareness about the harms of hate speech.

Problem definition Research question: What is the role of textual modality in hateful meme detection?

Sub-questions:

1) What are the advantages and disadvantages of different text classification models within the context of hateful meme detection?

2) What is the impact/role of stylometric and emotion-based features on/in the hateful meme detection?

Data-collection The dataset is already available at <https://www.kaggle.com/datasets/williamberrios/hateful-memes>. The dataset has been constructed by a professional team from FacebookAI and its quality has been checked and refined to make it very interesting to the task at hand.

research method During the project the following methods will be compared and explored:

- Textual models:

* 1. SVM with bag-of-words (baseline)
  2. SVM advanced using stylometric and emotion-based features [[1]](#footnote-2) [[2]](#footnote-3)
  3. Bert

The image classifier would be frozen throughout the whole experiment since we are focusing on the textual modality in the memes. A state-of-the-art image classifier will be chosen (to be determined later on).

To answer the second subquestion one of two approaches will be adapted:

1- performing a study where we add the features one by one to investigate the impact of each feature on the performance of the model.

2- performing a study comparing the gold labels to the features directly and trying to find patterns. This has the advantage that we eliminate other factors affecting the results (such as the model of choice)

|  |  |  |
| --- | --- | --- |
| Provisional organization of chapters | | * Abstract * Introduction * Related work * Methodology   + Data   + Experiment Setup * Results * Discussion   + Error analysis * Conclusion & Future Remarks |
| Provisional book list / intial references  (appendix) | * Hateful Memes Challenge and dataset for research on harmful multimodal content. (2020, May 12). ai.facebook.com. Retrieved April 3, 2023, from <https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/> * Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020b). The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. ArXiv (Cornell University). https://arxiv.org/pdf/2005.04790.pdf * Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C. A., Bull, P., Lipstein, G., Nelli, T., Zhu, R. X., Muennighoff, N., Velioglu, R., Rose, J., Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., . . . Parikh, D. (2021). The Hateful Memes Challenge: Competition Report. NeurIPS 2020 Competition and Demonstration Track, 344–360. * SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification * Gevers, I., Markov, I., & Daelemans, W. (2022). Linguistic Analysis of Toxic Language on Social Media. Computational Linguistics in the Netherlands Journal, 12, 33–48. Retrieved from <https://clinjournal.org/clinj/article/view/146> | |

**Timetable:**

|  |  |  |
| --- | --- | --- |
| **Phase** | **Dates** | **Tasks** |
| **orientation phase** | 6/2 – 31/3 | - Explore dataset and literature  - Determine the topic and setup the thesis planning  - look up features that could be used |
| **execution phase** | 31/3 – 14/3 | - writing related work  - finalize features to be extracted (theoretically)  - implementation of baseline |
| 14/3 – 28/4 | - implementation of complex model  - Bert |
| 28/4 – 12/5 | - finalizing code  - writing methodology |
| 12/5 – 26/5 | - writing results  - writing discussion  - start error analysis |
| 26/5 – 9/6 | - finalize error analysis  - writing error analysis  - writing conclusion & introduction & abstract |
| 9/6 – 16/6 | - Finalizing report  - Cleaning up the code |
| **completion phase** | 16/6 – 23/6 | Preparing presentation |

|  |  |
| --- | --- |
| start date | 01/04/2023 |
| planned thesis presentation date | xx/06/2023 |
| planned graduation date | 30/07/2023 |

Remarks .............................................…....................................…

1. Markov, I., Ljubesic, N., Fišer, D., & Daelemans, W. (2021). Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection. Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 149–159. [↑](#footnote-ref-2)
2. Ilia Markov and Walter Daelemans. 2021. Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate. In Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 17–22, Online. Association for Computational Linguistics. [↑](#footnote-ref-3)