**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)/ Dr. 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 ???

Topic The Hateful Meme detection challenge, created by Facebook, 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 how could linguistic cues affect the performance of the ML models.

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) How do textual modlaity act in regards to late fusion and early fusion ?

2) Which is the best performing system in identifying textual modality for hateful meme detection?

3) What are the limitations of textual modality in the automated detection of hate speech in memes?

4) What factors influence the effectiveness of linguistic cues for automated detection of hate speech?

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 and the quality of it 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:

1. Late fusion of linguistic features against early fusion.
2. Linguistic features such as:
   1. hateful word per sentence rate
   2. adjective usage
   3. laughter
   4. sentiment analysis
   5. intent detection
   6. (maybe) named entity recognition: Person, Land, and ethnicity since those could possibly be referring to some kind of hate speech
   7. (maybe) word embeddings
   8. *(maybe)* ***gender detection (FairFace could be used)***
   9. *(maybe)* ***race detection (FairFace could be used)***
   10. ***object detection***

We have text features and image features.

text models:

* Baseline: SVM bag of words
* SVM with complex features out [ref]
* Bert

The image classifier would be frozen throughout the whole experiment since we are focusing on the textual modality in the memes.

First two baseline models will be implemented or **taken from literature** using a straight forward algorithm. One for late fusion and one for early fusion.

* Early fusion architecture sketch: word embeddings using XXX+ image rich representation using XXX  SVM  probability prediction
* Two classifiers one for Text and one for Image:  merge results through mean or some other rule-based technique

Then the same algorithms will be adapted to include the presented features.

|  |  |
| --- | --- |
| Provisional organization of chapters | * Abstract * Introduction * Related work * Methodology   + Data   + Experiment Setup * Results * Discussion   + Error analysis * Conclusion & Future Remarks |

Provisional book list

(appendix)

Timetable

|  |  |  |
| --- | --- | --- |
| Phase | Dates | Tasks |
| orientation phase | 6/2 – 17/3 | - Explore dataset and literature  - Determine the topic and setup the thesis planning  - look up features that could be used |
| execution phase | 17/3 – 31/3 | - writing related work  - finalize features to be extracted (theoretically) |
| 31/3 – 14/4 | - implementation of features |
| 14/4 – 28/4 | - early fusion models  - late fusion models |
| 28/4 – 12/5 | - word embedding (optional)  - 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 |
| completion phase | 9/6 – 16/6 | - Finalizing report  - Cleaning up the code |
| 16/6 – 23/6 | Preparing presentation |

start date 01/04/2023

planned thesis presentation date /06/2023

planned graduation date 30/07/2023

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