**1**

Chess games are noted in a uniform form like the PGN format in order to be able to perform analyses and evaluations at a later date. These data usually contain information about the players and the event as well as the exact move sequence of the chess game. In addition, especially in large chess databases, decisive moves and positions are provided with comments after the game has been analyzed by grandmasters. For a better arrangement of these comments there are standardized symbols and NAGs, which for example directly indicate whether a move was good or bad. These have the advantage that they are easier to evaluate due to their clear categorization and are also generally understandable, which is not the case with comments in natural language. \\\\

In many cases the comments were therefore replaced by the annotation symbols if a comment did not seem necessary. For more detailed explanations, which can not only be expressed with an annotation symbol, both annotation symbols and comments are used. Finally, there are comments without a corresponding annotation symbol. This may be intentional if the commentary only names the commentator, summarizes a game or refers to similar chess games. However, some comments refer specifically to moves and positions and could therefore be provided with an annotation symbol. \\\\

For the evaluation of annotation symbols in chess there are already approaches as described in \autocite{Wirth2014} and \autocite{Wirth2015}, where the annotations are used to learn preferences. In order to be able to use pure comment data for such evaluations, annotation symbols must have been assigned to the comments beforehand.

Problem Definition

In this thesis, a suitable approach for the correct assignment of annotation symbols to given chess annotations is to be found. It is a specialization of a supervised sentiment classification problem since the used data for the learning process contains only comments that are already labeled with the correct annotation symbol. This problem is split up into five separate questions:

\begin{itemize}

\item the correct categorization into move- and position-annotating comments for given chess comments

\item the correct assignment of two annotation symbols $(!, ?)$ to given chess move comments

\item the correct assignment of six annotation symbols $(!!, !, !?, ?!, ?, ??)$ to given chess move comments

\item the correct assignment of three annotation symbols $(+-, =, -+)$ to given chess position comments

\item the correct assignment of seven symbols $(+-, \pm, +=, =, =+, \mp, -+)$ to given chess position comments

\end{itemize}

Goal of the Thesis

The main question of this work is whether and how well the appropriate annotation symbol for a given chess comment can be determined. Thus, the goal to find the configuration with the best accuracy and other evaluation metrics with regard to the classification problems just presented. For this purpose, four data models will be developed, ten classifiers selected and the configurations tested on three different data sets. In addition, further suitable processing methods are to be investigated and, if appropriate, added in order to optimize the results. The final analysis is also intended to provide insights into the remaining possibilities for improvement.

Structure of the Thesis

The thesis is split into five more sections: \\\\

The second section initially provides background knowledge in the topics dealt with in this thesis. It starts with the presentation of sentiment analysis as part of text mining. Afterwards the concept of word embeddings is discussed, which can serve as a data model in such text mining problems. This is followed by a description of multiclass classification problems and several suitable approaches to solve such problems. A separate look is taken at the subgroup of ordinal classification problems. Finally, cost-sensitive methods are presented that take into account different weightings of misclassifications. \\\\

The third section describes the general concept with which a text mining process can be created. It first deals with the requirements of problem and goal definition and the criteria for a suitable data selection. For the preparation of the data and their transformation into a model suitable for analysis, methods of natural language processing are presented. Finally, possible analysis methods and evaluation methods are presented from which suitable techniques for the problem can be selected. \\\\

In the fourth section, the procedure presented in the third section is applied to the text mining process in chess annotations. First the format PGN and the annotation symbols NAG are explained and the five problems are specified. In the following the used tools like NLTK, Weka and further libraries are listed. Finally, the classifiers and evaluation methods used are mentioned. \\\\

The fifth section contains all results and their evaluations. After the analysis of different tokenizer configurations, statistics about the comments are generated, which can be used to gain basic knowledge about the analyzed data set. In addition, the attributes and models used for the data set are evaluated. The majority of the results are finally taken up by the comparisons of the achieved accuracies for all configurations and problems. In the end, the best results achieved are checked for optimality in a cost-sensitive evaluation. \\\\

The last chapter summarizes the most important conclusions from the results and gives an outlook which further approaches could be used to improve the results.