HarryMotions – Classifying Relationships in Harry Potter based on Emotion Analysis

Albin Zehe **Julia Arns** **Lena Hettinger** **Andreas Hotho**
Data Science Chair — University of Würzburg
[zehe,arns,hettinger,hotho]@informatik.uni-wuerzburg.de

Abstract

Sentiment Analysis has long been a topic of interest in natural language processing and computational literary studies, where it can be used to infer the relationships between fictional characters. Building on the dataset and results of Kim and Klinger (2019), we propose a classifier based on BERT that improves the results reported therein and show that we can use this classifier to determine the relation between characters in Harry Potter novels. Our proposed sentiment classifier yields an F1-score of up to 75 % for binary classification of emotions. Aggregating these emotions over novels, we reach an F1-score of up to 68 % for the classification of a pair of characters as friendly or unfriendly.

1 Introduction

Characters and their relations are one of the basic building blocks of stories (Hettinger et al., 2015). Detecting them automatically is therefore a highly interesting task for the analysis of fictional texts. While there exists a multitude of methods for the extraction of character networks (Labatut and Bost, 2019), these often provide networks with unlabelled edges, that is, no information about the kind of relationship the characters share. Following Kim and Klinger (2019), we work towards the goal of detecting the polarity of relations using sentiment analysis. To this end, we collect all chunks of text in a novel mentioning a pair of characters and perform sentiment analysis on these pieces of text. While methods for sentiment analysis perform very well for certain domains, mostly short texts like tweets, product reviews or news articles, the task still poses a significant challenge on other domains. Fictional literary texts in particular are hard to analyse, since they usually do not express emotions explicitly, but they have to be inferred from context and possibly world knowledge.

Recently, the trend in NLP has been to use large transformer models that have been pre-trained for language modelling (or similar tasks not requiring explicit annotations) on enormous datasets. We follow this trend by fine-tuning BERT (Devlin et al., 2019) to the task of classifying emotions in interactions between characters. We use BookNLP (Bamman et al., 2014) to extract entity mentions and co-references and then fine-tune BERT on the emotion dataset provided by Kim and Klinger (2019). Emotions are aggregated to detect overall relations between characters and their development over a novel, as exemplified in Figure 1 (cf. Section 4).

Our contribution is two-fold: 1. We generally improve results on the emotion classification tasks from Kim and Klinger (2019). 2. We track the emotional relations detected by our classifier over the course of a novel and describe an easy method to aggregate them to an overall label. We evaluate this method on the text of the well-known Harry Potter series (Rowling, 1997).

The remainder of this paper is structured as follows: After giving a short introduction, we next present related work. In chapter 3 we describe our approaches towards emotion and relation classification as well as our results. We conclude this paper with a discussion of results and some possible directions for future work

2 Related Work

Our work is situated at the intersection of sentiment analysis and social network extraction.

Character networks for works of fiction have

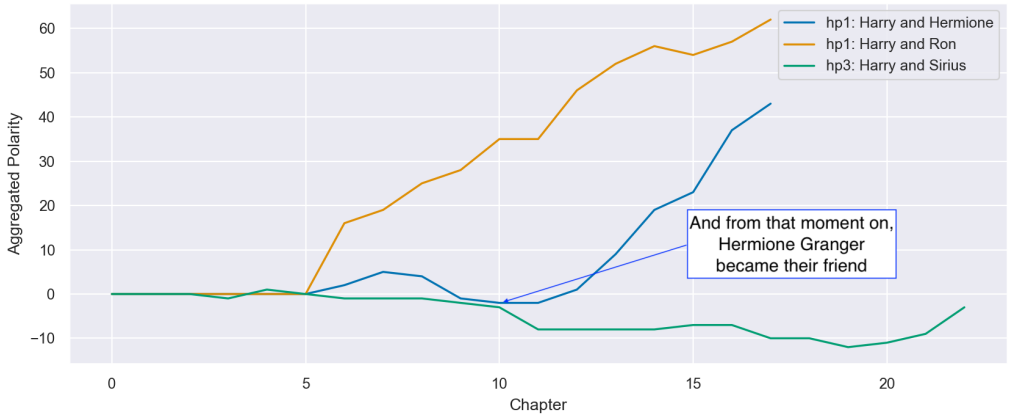

Figure 1: Trajectory of emotions for different character pairs in Harry Potter as detected by our system. The points where Harry and Ron/Hermione become friends are clearly visible. Details are discussed in Section 4. The x axis corresponds to chapters in the books, with book 3 having more chapters than book 1 and thus a longer trajectory.

been studied extensively in recent years (Labatut and Bost, 2019). Some work has been done on extracting networks from textual summaries (Chaturvedi et al., 2016; Srivastava et al., 2016) and training large neural networks to specifically model relationships over time (Iyyer et al., 2016). While Harry Potter novels have been explored before (Vilares and Gómez-Rodrıguez, 2019; Everton et al., 2019), research has not yet concentrated on emotional relations between characters.

For sentiment analysis, most work has focused on short, self-contained texts like tweets (Islam et al., 2019; Rosenthal et al., 2017) or reviews (Maas et al., 2011; Xue et al., 2020; Socher et al., 2013). Sentiment analysis in fictional texts has become a topic of interest, but has so far proven difficult because of the lack of suitable datasets. Kim and Klinger (2018) provide an extensive overview of papers addressing the issue of sentiment analysis in fictional texts, also addressing papers that use emotions in the context of social network extraction.

However, most of these works employ rather simple sentiment analysis methods (e.g., Zehe et al. (2016) rely on a simple lookup in a sentiment lexicon). Most similar to our work is Kim and Klinger (2019), which we directly build upon. The authors propose a new corpus of short pieces of text annotated with the emotional relations between characters described in these texts. They train a GRU (Cho et al., 2014) neural network to predict the emotions based on this corpus, showing promising results with F1 scores up to 67 % for undirected binary classification (positive and negative emotions) and 46 % for 5 basic emotions in the story-level evaluation as described below. We extend this work by improving the sentiment analysis model and aggregate the instance-level labels for full novels.

3 Classifying Emotional Relations

We address two tasks in this paper: mention-level *emotion classification* and story-level *relation classification*, which we see as two steps in a pipeline.

Emotion Classification Following Kim and Klinger (2019), we define emotion classification as learning a classifier that, given a short piece of text (roughly one sentence) containing two characters, predicts the emotion described therein. We perform this task on different granularity levels, using either 2, 5 or 8 directed or undirected emotions.

Relation Classification We define relation classification as an aggregation of emotions discovered by step 1 over a novel. In this paper, we distinguish between "friendly" and "unfriendly" relations.

3.1 Method

Emotion Classification We use a pretrained BERT-model (Devlin et al., 2019), which we fine-tune to our task using the fast-bert library[1], mostly keeping the default parameters. We train for 6 (2-, 5-class) or 12 (8-class) epochs with batch size 1.

Relation Classification We extract all interactions from a novel mentioning a pair of characters

[1] https://github.com/kaushaltrivedi/fast-bert, based on https://github.com/huggingface/transformers

a, b, classify the emotions described therein and aggregate them to an overall label. We use BookNLP (Bamman et al., 2014) to perform co-reference resolution and extract all interactions where both a and b each appear at least 20 times in the novel. We define an interaction as a chunk of text where a and b appear with no more than 10 tokens between them, regardless of sentence boundaries, with 10 additional tokens on both sides as context. We select only pairs where at least 5 interactions occur in the novel and classify the emotions in each of these interactions using our BERT-based classifier. For the aggregation of emotions to an overall relation, we count the number of positive, negative, neutral and overall emotions ($\mathcal{X}_{a,b}$) between a and b and calculate their difference, classifying relations as

$$\text{rel}(a,b) = \begin{cases} \text{friendly} & \text{if } \alpha < \frac{pos_{a,b}}{all_{a,b}} \\ \text{unfriendly} & \text{if } \alpha \geq \frac{pos_{a,b}}{all_{a,b}}. \end{cases}$$

The amount α of positive emotions required for a friendly relationship is a hyper-parameter.

3.2 Datasets

Emotion Classification For the first task, we use the dataset provided by Kim and Klinger (2019) and refer to this paper for a detailed description due to space constraints. The dataset consists of 1335 samples[2], each annotated according to multiple schemes. These schemes differ in the number of emotions that are annotated (two, five or eight) and whether the emotions are directed (from a causing to an experiencing character) or undirected.

Relation Classification For the second task, we have collected our own dataset. To this end, we used BookNLP on all books from the Harry Potter series to extract all interactions as described in Section 3.1. In contrast to the first dataset, we use automatically extracted characters and co-references here. We then manually annotated all pairs of characters for which we found interactions with their relationship, distinguishing between friendly and unfriendly relationships. We collected two sets of independent annotations and, where the two annotators disagreed, collected a third annotation as a tie-breaker. The tie-breaker was given the option to note that there is no (clear) relation between the two characters. This was the case in the third novel for the relation between Harry and Sirius Black (cf. Section 4). Table 1 provides details for the resulting dataset, which we publish for future research.[3]

Novel	#friendly	#unfriendly	#disagree
HP1	64	30	2/0
HP2	61	29	3/0
HP3	62	26	7/4
HP4	233	36	22/0
HP5	144	57	19/0
HP6	107	38	18/0
HP7	115	44	27/0

Table 1: Character relations in Harry Potter. Middle columns show friendly and unfriendly relations, respectively. Last column shows relations where a tie-breaker was used/no agreement could be reached.

3.3 Evaluation

Emotion Classification We follow the evaluation setup from Kim and Klinger (2019) for emotion classification, who use multiple settings: The dataset (cf. section 3.2) provides annotations for sets of two, five and eight directed or undirected emotions. Additionally, they define different ways of representing the entities involved in the emotions, where some add a marker to entities or completely mask them (making it impossible for the model to learn that, e.g., Harry always interacts positively with Ron). We describe these schemes shortly in the following and give an example for how sentences would be represented according to each scheme:

- **No-indicator**: Entities are represented as in the text, the model is directly fed the unmodified sentence (e.g., `Alice is angry with Bob`).
- **Role**: Entities are marked as causing or experiencing (e.g., `<e>Alice</e> is angry with <c>Bob</c>`), where `<e>` marks the experiencing character and `<c>` the causing character.
- **MRole**: Entities are only identified by their role (`<e> is angry with <c>`), `<e>` and `<c>` as above.
- **Entity**: Entities are marked as entities with no indication as to whether they cause or experi-

[2] 1742 overall, but following Kim and Klinger (2019) we use only the subset annotated with a causing character

[3] http://professor-x.de/datasets/harrymotions.

ence the emotion (e.g., `<et>Alice</et> is angry with <et>Bob</et>`).

- **MEntity**: Entities are masked by entity-markers (e.g., `<et> is angry with <et>`).

Table 2 shows our results in comparison to those from Kim and Klinger (2019), reporting what they define as story-level F1 score. Our classifier outperforms theirs in most settings, as discussed in Section 4.

Relation Classification In our second experiment, we use the emotions detected in the previous step to detect overall relationships between characters in the Harry Potter series by aggregating over emotions as described in Section 3.1. In Table 3, we report macro-averaged F1-scores as well as accuracies for aggregating emotions as classified in the Entity and MEntity settings for 2 and 5 emotion classes, since we do not have role labels for the Harry Potter corpus and the emotion classification for 8 emotions did not perform well. Note that the number of emotions only pertains to the emotion classification setting, relations are always classified as friendly or unfriendly. For the 5 class setting, we define anger, disgust and sadness as negative emotions, joy as positive and anticipation as neutral. The parameter α was optimised on hp1 and is set to 0.4, except for 5-MEntity ($\alpha = 0.75$). Lacking a directly comparable approach, we report sampling from the true label distribution per novel as a baseline (which performs better than majority vote in our setting). We find that, on average, 2 classes lead to better results and we always outperform the baseline.

4 Discussion

In this section, we discuss our findings along with some of the decisions involved in the dataset collection and provide some insight regarding the development of emotions over the course of a novel.

BERT vs. GRU Our BERT-based classifier outperforms the GRU in all undirected, but not all directed settings. Specifically, in the 8-class directed evaluation, the GRU usually performs better than BERT. We hypothesise two possible reasons: a) the rather low amount of training data available for each of the 8 emotion classes, especially in the directed case. We assume that the GRU's lower number of parameters makes it easier to tune on fewer samples. b) BERT is a bi-directional model, while the GRU used here is uni-directional. Since the GRU reads sentences in the right order, while BERT reads in both directions, it might be easier for the GRU to model directed relations.

Dataset Collection As mentioned in Section 3.2, we excluded some relations during the annotation process. This is due to two reasons: a) errors in named entity recognition and b) changing relationships. For the first category, BookNLP returned the entity "Felix Felicis", which is a luck potion. We excluded all relationships involving the potion, but kept collective entities like "Hogwarts". In the second category we find the relationship between Sirius Black and most other characters in the third novel. For the majority of the book, Sirius is regarded as a villain intent on killing Harry, which is revealed to be wrong at the end of the novel, turning the relation very positive. Since the label here is unclear, we excluded it from the dataset.

Developing Relations As described before, relationships can change drastically within a novel. Two prominent examples of this in the Harry Potter novels are the relations between Harry and Hermione in the first novel (where they become friends) and between Harry and Sirius Black in the third novel (see prev. paragraph). We can use the emotions detected by our classifier to plot a trajectory over the novel. The polarity for characters a and b in chapter i is then calculated as $p_i = p_{i-1} + pos_{a,b,i} - neg_{a,b,i}$, where $pos/neg_{a,b,i}$ counts positive/negative emotions between a and b in chapter i, respectively, and $p_0 := 0$. We show plots for three examples in Figure 1, using predictions from the 2-class MEntity classification. In all cases, the trajectory matches our expectation: For Harry and Hermione, the relation starts neutral with a very clear upper trend after they become friends. For Ron, the relation quickly becomes very positive. For Sirius, the relation is mostly negative, while improving clearly in the final chapters.

5 Conclusion

We have presented an improved approach for the classification of emotional relations between fictional characters. By aggregating sentence level emotions, we have built a classifier for novel-wide character relations based on emotion analysis. While our experiments show that aggregation yields promising results, future work includes

	GRU						BERT					
	Undirected			Directed			Undirected			Directed		
Setting	8c	5c	2c	8c	5c	2c	8c	5c	2c	8c	5c	2c
NoInd	33	41	66	**25**	23	37	**34**	**52**	**74**	21	**29**	**41**
Role	**19**	34	55	**33**	**35**	**56**	**19**	**51**	**65**	21	23	34
MRole	32	44	67	**39**	44	65	**39**	**59**	**75**	30	**55**	**75**
Entity	21	31	57	**22**	18	30	**28**	**44**	**70**	18	**30**	**46**
MEntity	33	46	65	28	30	39	**34**	**55**	**74**	**31**	**36**	**48**

Table 2: Comparison of story-avg F1-scores between our classifier (BERT) and the GRU from Kim and Klinger (2019). Results for the GRU are taken from the original paper. The best result for each setting is marked in **bold**.

	F1-score					Accuracy			
	2-class		5-class			2-class		5-class	
Novel	En	MEn	En	MEn	Base	En	MEn	En	MEn
hp1*	53	61	**64**	60	46	55	64	**69**	61
hp2	**57**	56	37	52	41	**62**	59	38	56
hp3	62	60	**68**	56	45	64	61	**70**	56
hp4	60	**68**	56	55	60	72	**77**	67	64
hp5	**62**	56	57	**62**	47	64	58	61	**65**
hp6	**60**	58	**60**	**60**	46	62	62	**63**	63
hp7	58	57	**63**	49	46	62	60	**68**	50
avg	**59**	**59**	58	56	47	**63**	**63**	62	59

Table 3: Macro-averaged F1-scores and accuracies for the classification of relations according to different emotion annotation schemes. *En* refers to the *Entity* annotation scheme, *MEn* to the *MEntity* scheme. Base refers to the stratified random baseline. hp1* was used as a development set to determine the value of α.

the development of a stronger classifier for story-level relations. We also plan on investigating the influence of co-reference resolution, which is currently done automatically. Using manual labels or improved co-references resolution should further improve our results: First experiments indicate better performance for frequent characters, where resolution errors are more easily smoothed out.

Acknowledgements

Many thanks to Darleen Pappelau for helpfully providing the tie breaker annotations for the dataset.

References

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling dynamic relationships between characters in literary novels. *Thirtieth AAAI Conference on Artificial Intelligence*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sean Everton, Tara Everton, Aaron Green, Cassie Hamblin, and Rob Schroeder. 2019. Strong ties and where to find them: Or, why Neville (and Ginny and Seamus) and Bellatrix (and Lucius) might be more important than Harry and Tom. *SSRN*.

Lena Hettinger, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2015. Genre classification on german novels. In *Proceedings of the 12th International Workshop on Text-based Information Retrieval*.

Jumayel Islam, Robert E. Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1355–1365, Minneapolis, Minnesota.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. Submitted for review to DHQ (http://www.digitalhumanities.org/dhq/).

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

J. K. Rowling. 1997. *Harry Potter and the Philosopher's Stone*, 1 edition, volume 1. Bloomsbury Publishing, London.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Thirtieth AAAI Conference on Artificial Intelligence*.

David Vilares and Carlos Gómez-Rodrıguez. 2019. Harry Potter and the action prediction challenge from natural language. In *Proceedings of NAACL-HLT*, pages 2124–2130.

Qianming Xue, Wei Zhang, and Hongyuan Zha. 2020. Improving domain-adapted sentiment classification by deep adversarial mutual learning. Accepted to appear in AAAI'20.

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. Prediction of happy endings in german novels. In *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing 2016*, pages 9–16.