DigitCult | Scientific Journal on Digital Cultures

# A New Research Programme for Reading Research: Analysing Comments in the Margins on Wattpad

Simone Rebora
Göttingen Centre for Digital Humanities
University of Göttingen
Papendiek, 16 – 37073 Göttingen (Germany)

Federico Pianzola
University of Milan Bicocca
Piazza dell'Ateneo Nuovo, 1
20126 Milano (Italy)

## Abstract

This paper focuses on Wattpad, a social reading platform on which people can add comments in the margins of books. Analysing these comments enables the comparison between specific parts of the text and the effects they have on readers. We outline a new research programme, discussing both theoretical and practical issues in the study of Wattpad: from the identification of a methodology holding together reader response theory, cognitive literary studies, and computational text analysis, to the definition of a digital mixed method for the recognition of the linguistic and textual cues that trigger certain effects. We describe a dataset built by scraping the Wattpad website: preliminary statistics on the most commented books in the categories "Classics" and "Teen Fiction" are presented and discussed. To provide an example of the possible uses of the dataset, we introduce a simplified experiment with the sentiment analysis software *Syuzhet*. By comparing the "emotional arcs" produced in parallel by text and comments, we evaluate the approach and show the substantial differences between the intrinsic emotional valence of the text and the effects it produces.

*Published* 29 September 2018

Correspondence should be addressed to Simone Rebora, Göttingen Centre for Digital Humanities, University of Göttingen, Papendiek, 16 – 37073 Göttingen (Germany). Email: simone.rebora@phil.uni-goettingen.de

*DigitCult, Scientific Journal on Digital Cultures* is an academic journal of international scope, peer-reviewed and open access, aiming to value international research and to present current debate on digital culture, technological innovation and social change. ISSN: 2531-5994. URL: http://www.digitcult.it

 doi:10.4399/97888255181532

## Introduction

In the 21st century, reading literature has become a widely diversified phenomenon: in fact, nowadays it includes many different practices that go beyond traditional processes involving only authors, solitary readers, publishers and their distributors (Graham and Gandini 2017). In particular, digital media have widened the possibilities available to authors and readers, often enabling them to establish a direct contact and to oust publishers, as it happens with the spreading phenomena of self-publishing (Dilevko and Dali 2006) and social reading (Cordón-García et al. 2013).

It is becoming increasingly urgent to consider and analyse how reading practices are changing, in order to outline a more reliable scenario of what kind of readers exists nowadays, how much they read and what channels they use to approach literature. Unfortunately, publishers are not fully aware of the magnitude of the reading practices that are out of their control, for instance, omitting to take into account data about self-published books in their reports about general book sales (Hoffelder 2017). Likewise, scholars are investigating only a limited range of aspects of the various emerging phenomena related to creating, publishing and reading literature. In this paper, we want to point out a new way of analysing online social reading – one of the most interesting phenomena related to the consumption of literature – showing the potential benefits of this research program. We are focusing on an analysis of the comments in the margins of some books on the social reading platform Wattpad.

## Comments in the Margins and the Social Reading Landscape

"Social reading" is a term encompassing a wide variety of practices mainly related to the activity of reading and using social media to talk about the reading experience. Traditional book clubs are a form of social reading too (Williams 2017), but nowadays the term is used almost exclusively in relation to the use of digital and social media. In this regard, studying how people share their reading experiences online is interesting to evaluate "how digital media are creating new social valences of reading" (Nakamura 2013, 238). Cordón-García et al. (2013, 156) in their recognition of the social reading landscape claim: "We understand 'social reading' to mean reading carried out on virtual environments where the book and the reading favour the formation of a 'community' and a means of exchange." This is only a partial definition that excludes practices like the sharing of annotations and highlights, which arguably lead to the formation of a community. However, it is true that the community is a key element in the most interesting and long-lasting online social reading phenomena, like Wattpad. Other authors have interestingly pointed out that the concept of "social reading" is inaccurate in two ways: on one hand, reading as a social practice is something that extends beyond virtual environments; on the other hand, "the practices to which the concept refers include more than just reading, e.g. also writing, distributing, criticizing, adapting, etc." (Vlieghe, Muls, and Rutten 2016, 27).

In this paper we are focusing on the kind of social reading called "discussion in the margin" (Stein 2010), which includes the quite famous The Golden Notebook Project, a website hosting the altogether formal conversation of seven authors commenting Doris Lessing's novel. But the category also includes most of the writing happening on Wattpad, a platform connecting writers and readers, where many comments are definitely informal and slangy.

## Wattpad

Wattpad is a very important resource for everybody interested in literature (cf. Miller 2015). It is a platform available via web and as a mobile app, on which people can add comments in the margins of books in the public domain, writing their own response to what they are reading and engaging in discussions with other users that commented before them (as shown in Figure 1).

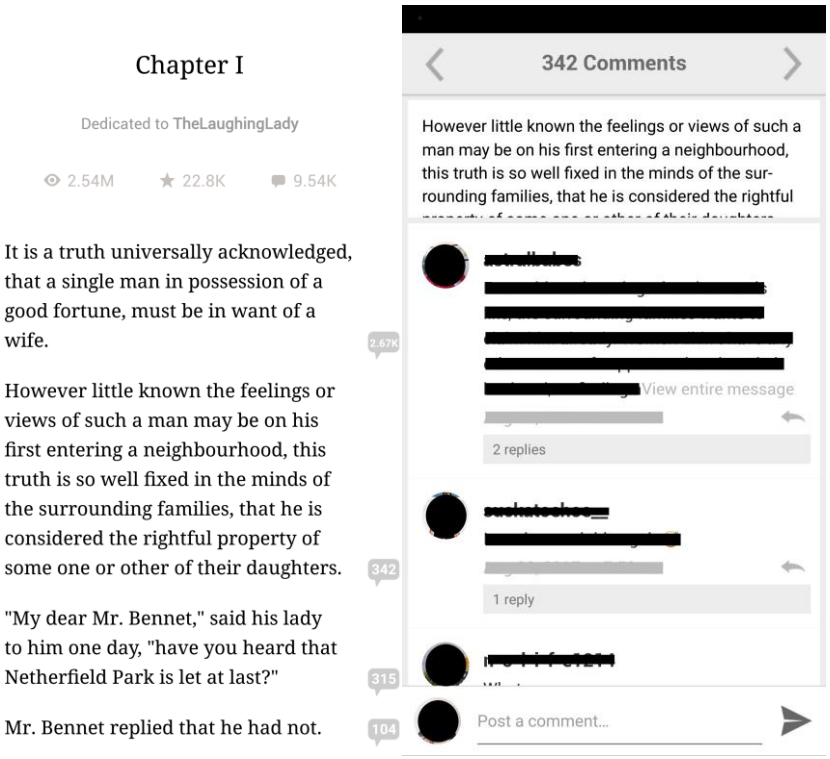


**Figure 1.** Two screenshots showing Wattpad's user interface for smartphones. On the left there is the reading interface, on the right the interface opening after a tap on the balloon with the number of comments. In the commenting interface: on the top there is the paragraph from the novel that is being read, below there are the users' comments (usernames and comments have been redacted for privacy reasons).

A few scholars have started exploring what is published on the platform (Mirmohamadi 2014; Fast, Vachovsky, and Bernstein 2016), and the dynamics between readers and writers using it (Ramdarshan Bold 2016). However, we think that one of Wattpad's most remarkable – but so far also most neglected – aspects is that the users' comments are generated *during* the reading activity. This is the most striking difference with respect to other social reading practices, since the social aspect of social reading – i.e. the production of user generated contents – is usually something that happens once reading a book or a story is concluded, like in the case of writing reviews, rating and recommending books, or organizing one's own online bookshelf on Goodreads.

## A New Research Programme

The difference is remarkable, because reader response changes and is shaped by the progression of reading. What the reader might think or feel in relation to the first chapter of a book can be drastically reshaped and reconfigured when reading the following chapters. Therefore, in contrast to a review, the comment in the margin can offer a "real-time" insight into the reading experience. From a reading research perspective, we regard this as the most valuable feature of the social reading happening on Wattpad, since it enables the collection of a kind of data that so far has been unavailable. More specifically, analysing the comments in the margins enables the comparison between a specific part of the text and the effects it has on readers. Not just a few readers, but millions of readers, in some cases (see the section "Scraping Wattpad" below).

Like any new research programme, this kind of research faces some critical issues that concern both the analysis of the data and the underlying assumptions guiding their interpretation. We started to reflect on how to design a useful and effective research methodology based on a sound epistemological ground, and this paper is just a first case study to test some preliminary ideas and their application. For the sake of clarity, we grouped the critical aspects of this research

 doi:10.4399/97888255181532

programme into "theoretical and methodological issues" and "practical issues" that we think it is worth pointing out in order to better understand the intentions underlying our choices.

## Theoretical and Methodological Issues

The first step required when approaching a phenomenon is to clarify the goal of the research, because this requires making a basic epistemological choice that will underlie the whole research project. In the case of studying the comments in the margins, the fundamental question to answer is: are we interested in extending our knowledge regarding language uses in literary and narrative texts? Or are we interested in the effects and the impact that literature has on readers? Choosing either the former or the latter option does not exclude that we can indirectly learn something about the other aspect, but this is a very important epistemological assumption that orients the design of the research methodology and affects the choice of the tools that we will use for the analysis. In this moment, we are more interested in the second aspect: reader's response to literary narratives. Thus, we will first focus on the readers' comments and compare them with the portion of texts that triggered the response, exploring the relations between the two datasets. In this respect, our research attitude is closer to transactional reader response theory (Iser 1978; Rosenblatt 1978) and second generation cognitive literary studies (Kukkonen and Caracciolo 2014; Caracciolo 2014) – which are looking at the interdependence between textual cues and the readers' experience – but we acknowledge that the research programme we are sketching could also be oriented by subjective criticism (Bleich 1978), focusing only on readers' interpretations.

Since at the core of this research programme there is a comparison between two domains – the text (forms) and the comments (effects) – we also have to reflect on what kind of information and knowledge we can obtain about one domain by observing the other domain. We subscribe to the position claiming that there are not predetermined links between forms and effects, that is "a certain function can be accomplished by different discursive forms, and a certain form can accomplish many different functions" (Passalacqua and Pianzola 2016, 209–10; cf. Sternberg 1992). In brief, even though we will identify recurring patterns in the effects, it can be the case that they are triggered by different textual cues and, likewise, recurring textual patterns could trigger different readers' responses.

This position does not contradict the typical stance of empirical studies that an aesthetic phenomenon can be quantified when observed on a significantly wide portion of a population (van Peer, Hakemulder, and Zyngier 2012). With reference to sentiment analysis (see the paragraph "Wattpad as a New Resource for Sentiment Analysis" below for more details), it counters the idea beyond certain generalizations that are typical in quantitative (computational) methods – like the "emotional story arcs" identified by Jockers (Archer and Jockers 2016) and Reagan et al. (2016). Our approach privileges instead the interaction and the possible discrepancies between textual features and readers' responses, more in line with cognitive studies such as Jacobs et al. (2017).

Furthermore, considered the kind of data that we are going to analyse, we will also need to use computational methods in order to manage the extension of the dataset. These premises bring us to face some further issues.

## How to Design a Research Method Holding Together Reader Response Theory, Cognitive Literary Studies and Computational Text Analysis?

The greatest challenge is to manage two approaches that focus on the reader's cognitive and emotional processes with methods and tools that necessarily deal with more tangible linguistic data. A theoretical hypothesis that we can point out is to rely on some literary/narrative theories that we think are fit for the goal and can possibly be combined. For instance, one option is to use Meir Sternberg's narrative theory – focusing on the narrative effects of curiosity, suspense, and surprise (Sternberg 1992) – and Marco Caracciolo's theory of "narrative experientiality" – focusing specifically on the processes of "consciousness enactment" and "consciousness attribution" emerging in text-reader interaction (Caracciolo 2014). These two theoretical models are epistemologically consistent with each other (Pianzola 2017), therefore they can be applied in combination. They are helpful in this effort since they aim at describing the cognitive and aesthetic processes involved when we read, trying to grasp how textual forms participate in the emergence of narrative effects. However, they can grasp only some aspects of readers' response to literary

texts, therefore they will have to be complemented by other compatible models that focus on other kinds of aesthetic and social effects.

### How Do We Account for the Differences Between Comments Produced in a Social Reading Context and Personal Annotations?

Since the comments in the margins on Wattpad are produced within a social reading context, we need to consider that their content is different from that of private annotations. Both can be regarded as social reading practices, since personal notes written on an ebook reader or app can also be made public thanks to the sharing function available in many software (Rowberry 2016). However, people annotate books for many different reasons (Melanie Ramdarshan Bold and Wagstaff 2017) and sharing them can be perceived as not relevant in some uses of social reading applications (Li, Wu, and Wang 2017). Regarding comments in the margins on Wattpad, two crucial aspects concern the reading purpose – readers use Wattpad for pleasure, mainly to read fanfiction or emerging authors – and the expectation set by the most consumed genre on the platform, Teen Fiction, which has a serialized publishing system that affects the readers' activities and their interactions (Davies 2017, 52). These aspects create a context in which there is a colloquial dimension and a widespread use of slang and abbreviations typical of online chats and social networks (e.g. "U so smoooth paps", a comment about a character). Furthermore, a random exploration of Wattpad brought us to notice that many comments have a genuine social function, like questions addressed to other readers (e.g. "This was the mentality then, no? 1800s?"). All these aspects complicate the task of cleaning and refining the data, and also affect the methodology chosen to frame and interpret them, if we want to focus on the aesthetic reader response only. On the other hand, the great number of socially driven comments make Wattpad a very interesting resource to investigate the social function of reading and the social dynamics that bring to the emergence and negotiation of meanings and interpretations.

## Practical issues

Our main goal is to compare the text with the effects it has on readers, but in order to do that we have to consider a few practical aspects regarding the collection and analysis of data.

### How to Identify Linguistic and Textual Cues that Trigger Certain Effects on Readers?

This is a topic addressed very often by narratological research and literary theory (e.g. Rosenblatt 1978), as well as by stylistics and textual linguistics (e.g. Weinrich 2001). One way to do it is to rely on what is explicitly referred to in the comments: for instance, when someone writes "Weird how he called her handsome", we know that the word "handsome" in the original text triggered the reader response of thinking that the character is saying something weird. However, we can intuitively claim that the number of this kind of comments will be just a small part of the whole. In the following subsections we will sketch a few hypotheses, but there is a broader question to address first. Given the many different possible correlations between forms and effects, which are dependent on what is perceived by readers: can textual forms be traced with the help of automated processes or do they need to be manually identified by readers, according to their response? And in which way can we train a machine to detect possible matching between readers' responses and textual cues? This is an exciting question on which we are reflecting but it is not of primary concern for this preliminary study.

A possible way to approach the issue is the one followed in the ongoing SANTA project (Systematic Analysis of Narrative Texts through Annotation), whose goal is the collaborative creation of annotation guidelines for narrative levels and the narrator position. The guidelines will be subsequently used for the automatization of narrative analysis. Narratological concepts – like "heterodiegetic narrator" or "character focalization", for instance – are a formalization of the reader's perception of some effects triggered by certain uses of language. Therefore, creating annotation guidelines is a way to describe a certain kind of reader response.

Regardless of the method used, suppose that we were able to identify some textual features with a satisfying accuracy and that these results would have been validated by the shared

 doi:10.4399/97888255181532

agreement of many readers. At that point, an additional difficulty would be to find a meaningful way to link the results of the analysis of the literary text with the effects identified in the comments.

### A Broad and Complex Operative Hypothesis for the Comments-Text Comparison: A Digital Mixed Method

The first step will be to map the comments in order to understand what kind of thoughts readers share: are they about the plot, about style and language, about the direct emotional effects the text has on them, or about other mental connections triggered by the text? We have the advantage of being able to work in parallel on the literary text and on the comments, having a great quantity of readers' responses linked to single paragraphs. This condition would ideally allow us to develop a "digital mixed methods" approach (cf. Herrmann 2017) by applying a set of tools and techniques – like topic modelling (Blei and M. 2012) and word embedding (Mikolov et al. 2013), or lexical databases such as *WordNet* (Fellbaum 2010) – to create networks that could help classifying the different types of (semantic) connections between texts and comments. Along this line, at a later stage, it would be possible to train a machine learning algorithm to identify the types of comments – after some categories have been precisely defined following suggestions by readers and critics – and the textual cues that possibly triggered them.

### A Narrower Operative Hypothesis to Start Exploring the Phenomenon: Emotional Arcs

In order to start working as soon as possible on this rich and innovative corpus, we decided to design a simple experiment using a quite common technique: sentiment analysis. The goal of this study is to gain a first insight about the diversity of the two datasets – literary text and comments – and about their relationships. We compared the "emotional arcs" of the text and of the comments. This operation can be done on the totality of the comments but also for every single user, comparing different readers' responses to the text. We did it on the totality of the comments since this is a preliminary test to verify if a narrower operative hypothesis could lead to useful and satisfactory results. The outcome of the experiment will hopefully allow us to better understand the validity and the limits of the instrument when applied to comments in the margins, thus not focusing on the text but on reader response. Understanding how sentiment analysis performs differently when applied to comments will help us obtain information about which parts of the text elicit the most remarkable responses. Further and more in-depth analysis will be needed to understand the reasons of these responses: are they directly related to the text paragraph? How the discussion between readers affect the tone of the comments? Why, if this is the case, readers' response differs from plot emotional values? Similar questions will be addressed in later research. In this paper, we are more focused on learning how to explore and extract information from this new kind of corpus.

## Scraping Wattpad

Collecting data for the analysis of emotional arcs is a process that has important ethical implications, because Wattpad's terms of service clearly state that they do not want the website to be scraped. This is a way of protecting the commercial value of the service offered by Wattpad – from which the company profits hugely. Indeed, Wattpad's concern is related to "any use of the Site, content or Services that may have the effect of competing with or displacing the market for Wattpad, the Site, or the Services". We consulted directly with Wattpad and we decided to proceed with the scraping, since our intention is not to harm Wattpad or its users in any way. Our goal is to investigate a phenomenon that is widely popular and meaningful for our societies, pushing us to reconsider not only the tools we use for our investigation, but also the boundaries of literary studies. An additional issue regards the age of Wattpad's users: since many of them are underage, we are under the obligation of protecting all their sensitive data. Therefore, we redacted all the usernames and the content from the tables and figures to prevent any identification. For the same reason, we are not allowed to openly share the dataset we created.

In order to collect enough data for the analyses, we set up a complex process to scrape the Wattpad website. The web pages have a dynamic structure: most of the content requires a click to be displayed and it is not accessible at specified URLs nor visible in the source code of the

page. For this reason, we had to use a virtual browser activated by an algorithm coded to simulate different interactions with the website. For instance, to download the over 2,600 comments to the first paragraph of Pride and Prejudice the following operations are required:

1. click on the balloon icon that opens the dynamic windows with the comments;
2. click more than 100 times on the "show more" button to visualise more than 10 comments;
3. click on all the "reply" buttons to visualise the replies to the comments (and click on all the "show more" buttons when the replies are more than 10).

The algorithm has been written in R language and developed on the *Remote WebDriver Selenium 2.0* (based on the *Docker* platform). In order not to overload the Wattpad server, we set breaks between 1 and 3 seconds after each operation. Despite the huge number of clicks required, the algorithm is quite stable, although quite slow: it took around ten hours to download the over 42,000 comments to *Pride and Prejudice*[1]. The data collected have been structured as exemplified in Table 1, by associating each comment (and its metadata) to the corresponding chapter and paragraph.

**Table 1.** Sample of the Wattpad corpus. The column "Reply" indicates whether the comment is a reply to the previous one.

| | Book | Chapter | Paragraph | Username | Date | Comment | Reply |
|---|---|---|---|---|---|---|---|
| 1 | Pride and Prejudice (1813) | Chapter I | It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. | [redacted] | Jan 16, 2018 | [redacted] | FALSE |
| 2 | … | … | … | [redacted] | Jan 16, 2018 | [redacted] | FALSE |
| 3 | … | … | … | [redacted] | Jan 15, 2018 | [redacted] | FALSE |
| 4 | … | … | … | [redacted] | Jan 17, 2018 | [redacted] | TRUE |
| 5 | … | … | … | [redacted] | Jan 08, 2018 | [redacted] | FALSE |

## The First Statistical Data on Wattpad Readers

A first round of scraping has been done to collect the reading statistic for each book in the "Classics" and "Teen Fiction" categories: numbers of readings, votes, and comments. This step has been necessary to design the subsequent download of the comments. Indeed, Wattpad does not provide in any of its pages an overview of the most read or commented books: the pages "Hot" and "Stories" are not ordered according to these criteria. Once these statistics have been gathered, we were then able to select which books to start scraping for the comments. Table 2 shows a sample of the statistics collected.

[1] As a comparison: for the over 2.5 million comments to *The Bad Boy's Girl* (the most commented book in the "teen fiction" category), it took around 3 weeks.

 doi:10.4399/97888255181532

**Table 2.** Statistics for the "Classics" and "Teen Fiction" categories on Wattpad.

| | Classics | | | | | Teen Fiction | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Book | Total comments | Read count | Vote count | | Book | Total comments | Read count | Vote count |
| 1 | *Pride and Prejudice* | 42,013 | 7,400,000 | 113,000 | 1 | *The Bad Boy's Girl* | 2,569,405 | 197,000,000 | 3,400,000 |
| 2 | *Romeo and Juliet* | 11,607 | 3,100,000 | 36,700 | 2 | *I Sold Myself to the Devil for Vinyls... Pitiful I Know* | 2,052,682 | 92,900,000 | 2,000,000 |
| 3 | *Wuthering Heights* | 6,653 | 1,700,000 | 13,200 | 3 | *She's With Me* | 1,788,844 | 102,000,000 | 3,700,000 |
| 4 | *Jane Eyre* | 6,177 | 1,600,000 | 16,700 | 4 | *The Hoodie Girl* | 1,567,444 | 58,000,000 | 2,200,000 |
| 5 | *Alice's Adventures in Wonderland* | 3,261 | 1,100,000 | 11,100 | 5 | *The Last Virgin Standing* | 1,412,758 | 61,900,000 | 1,600,000 |
| 6 | *The Picture of Dorian Gray* | 2,768 | 1,000,000 | 8,800 | 6 | *My Brother's Best Friend* | 1,204,380 | 114,000,000 | 2,200,000 |
| 7 | *Emma* | 2,137 | 1,200,000 | 8,900 | 7 | *The Cell Phone Swap* | 1,118,017 | 100,000,000 | 2,100,000 |
| 8 | *Great Expectations* | 1,767 | 1,300,000 | 8,500 | 8 | *The Bad Boy, Cupid & Me* | 1,004,800 | 64,000,000 | 1,700,000 |
| 9 | *Little Women* | 1,636 | 498,000 | 9,300 | 9 | *Mr. Popular and I* | 843,820 | 99,000,000 | 1,700,000 |
| 10 | *Anna Karenina* | 1,595 | 1,100,000 | 16,700 | 10 | *My Wattpad Love* | 733,900 | 47,200,000 | 1,400,000 |
| 11 | *Dracula* | 1,546 | 290,000 | 4,800 | 11 | *Breaking The Bad Boy* | 721,200 | 29,100,000 | 978,000 |
| 12 | *Anne of Green Gables* | 1,255 | 389,000 | 9,800 | 12 | *Stay With Me* | 682,194 | 25,800,000 | 1,200,000 |
| 13 | *The Adventures of Sherlock Holmes* | 1,232 | 454,000 | 6,500 | 13 | *Bad Boy's Game* | 668,489 | 52,000,000 | 1,600,000 |
| 14 | *A Tale of Two Cities* | 1,034 | 300,000 | 3,400 | 14 | *Must Date The PLAYBOY!* | 661,865 | 100,000,000 | 1,700,000 |
| 15 | *Macbeth* | 1,021 | 125,000 | 1,900 | 15 | *Growing up (MWL's sequel)* | 659,900 | 23,500,000 | 760,000 |
| 16 | *The Importance of Being Earnest* | 975 | 134,000 | 1,800 | 16 | *The Quirky Tale of April Hale (Quirky Series #1)* | 637,304 | 43,200,000 | 1,100,000 |
| 17 | *A Midsummer Night's Dream* | 845 | 112,000 | 2,000 | 17 | *Silently Falling* | 608,528 | 24,300,000 | 1,100,000 |
| 18 | *Demian* | 769 | 79,600 | 1,400 | 18 | *The President's Daughter* | 569,000 | 42,900,000 | 1,100,000 |
| 19 | *Hamlet* | 757 | 140,000 | 2,000 | 19 | *Started With a Lie* | 554,976 | 49,800,000 | 1,000,000 |
| 20 | *Oliver Twist* | 719 | 280,000 | 4,100 | 20 | *Just A Friend?* | 554,208 | 36,300,000 | 1,000,000 |

We decided to focus on the "Classics" and "Teen Fiction" categories because the former is the most interesting in the broader context of literary studies, and the latter contains the most read and commented books of the whole website (as per our judgment, after browsing the website for a couple of weeks). There is a very clear difference between the two categories: the first 8 Teen Fiction books have over 1 million comments, whereas the most commented Classic book does not even reach 50,000 comments and the 16th is already under 1,000 comments. It can also be noted that books written by women (Jane Austen and the Brontë sisters) or with female protagonists (*Alice's Adventures in Wonderland*; *Anna Karenina*) are very popular and commented. This fact seems to renew the emphasis on gender issues in the Western literary canon (Winders 1991) and its global reception, since many Wattpad users are from non-Western countries, like the Philippines, Turkey, Mexico, and India (Miller 2015, 2). More accurate analyses about the Wattpad community will be required but even a simple classification of users by gender is quite hard to do, since many user accounts are incomplete or fictitious (for a similar problem cf. Thelwall and Kousha 2017, 975–76).

Anyway, we were able to determine some information from the data collected so far: 6,219 users engaged with *Pride and Prejudice* between December 2013 and January 2018 (5.9 comments per user on average), whereas *The Bad Boy's Girl* had 138,832 active users (18.7 comments per user on average) and attracted users more regularly in the time span considered (Figure 2). Another interesting information regards the variation of the number of comments over the progression of the book: in general, Teen Fiction shows more stability (Figure 3), whereas for many Classics the majority of comments is on the first chapter or the incipit (Figure 4), although there are some exceptions (Figure 5). Users can also link comments to the chapter, rather than to a single paragraph, but these comments are less interesting for our analysis because they are less informative for investigating "real time" reader's response. Therefore, we omitted them from our textual analyses, but not from the general reading statistics.

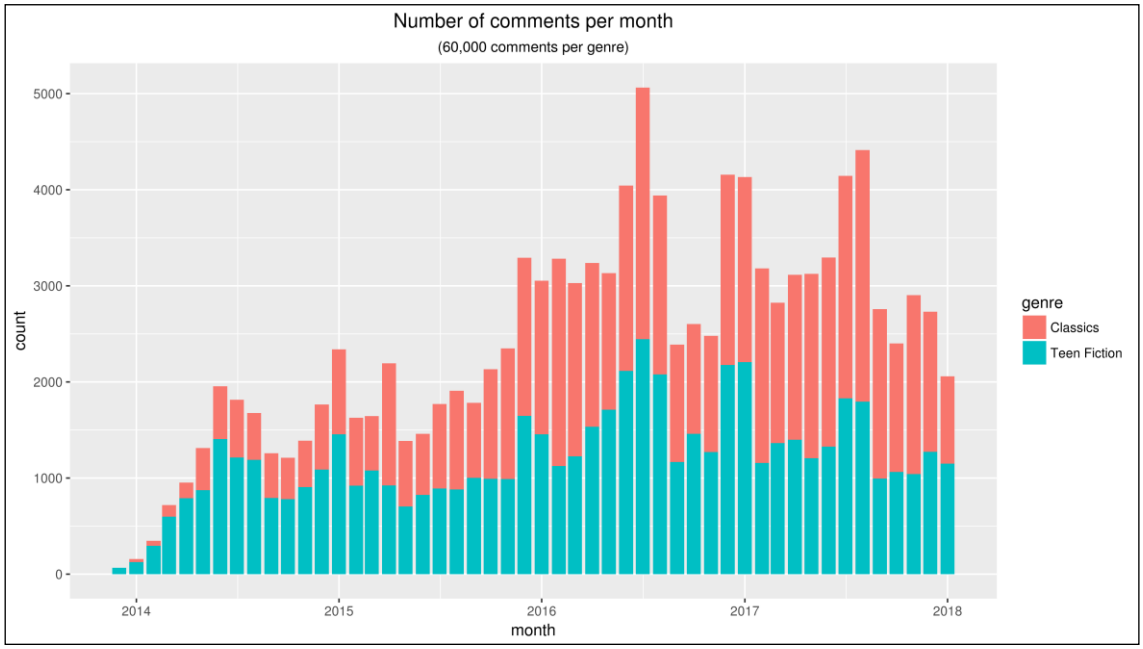


**Figure 2.** Number of Wattpad comments for the categories “Classics” and “Teen Fiction”, ordered by publication date. This graph shows the total number of comments per day, stacking the values of both categories, out of a sample of 120,000 comments.

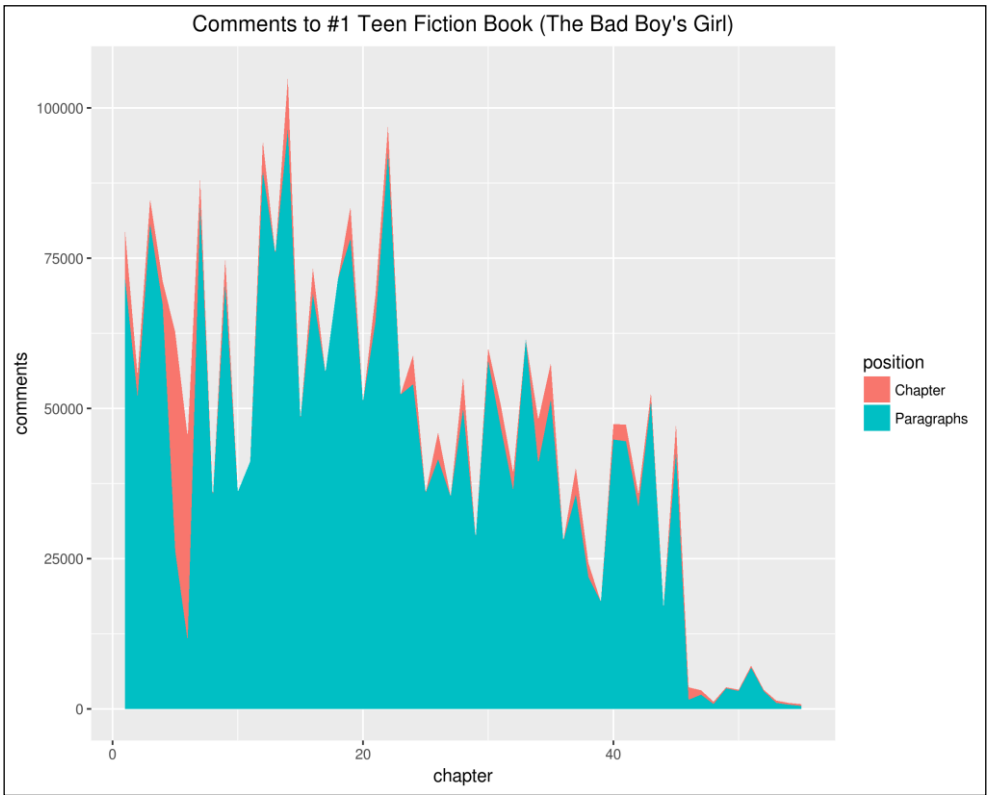


**Figure 3.** Number of Wattpad comments to *The Bad Boy’s Girl*, divided per chapter (chapters after XLV are an anticipation of the next book in the series).

 doi:10.4399/97888255181532

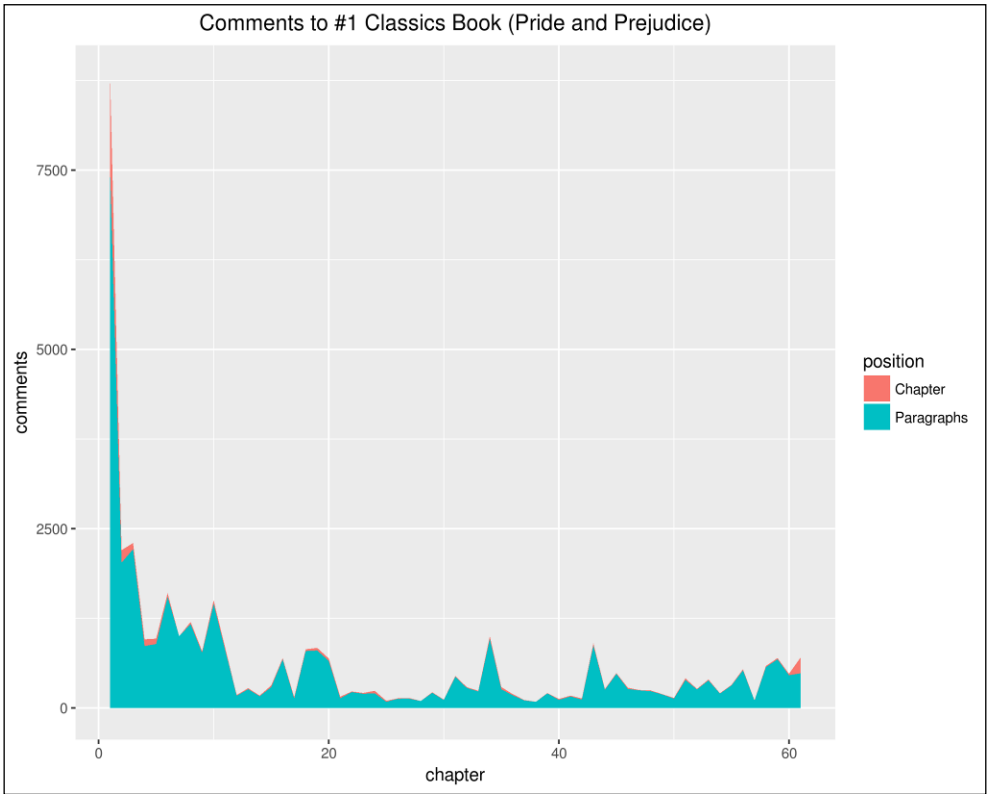


**Figure 4.** Number of Wattpad comments to *Pride and Prejudice*, divided per chapter.

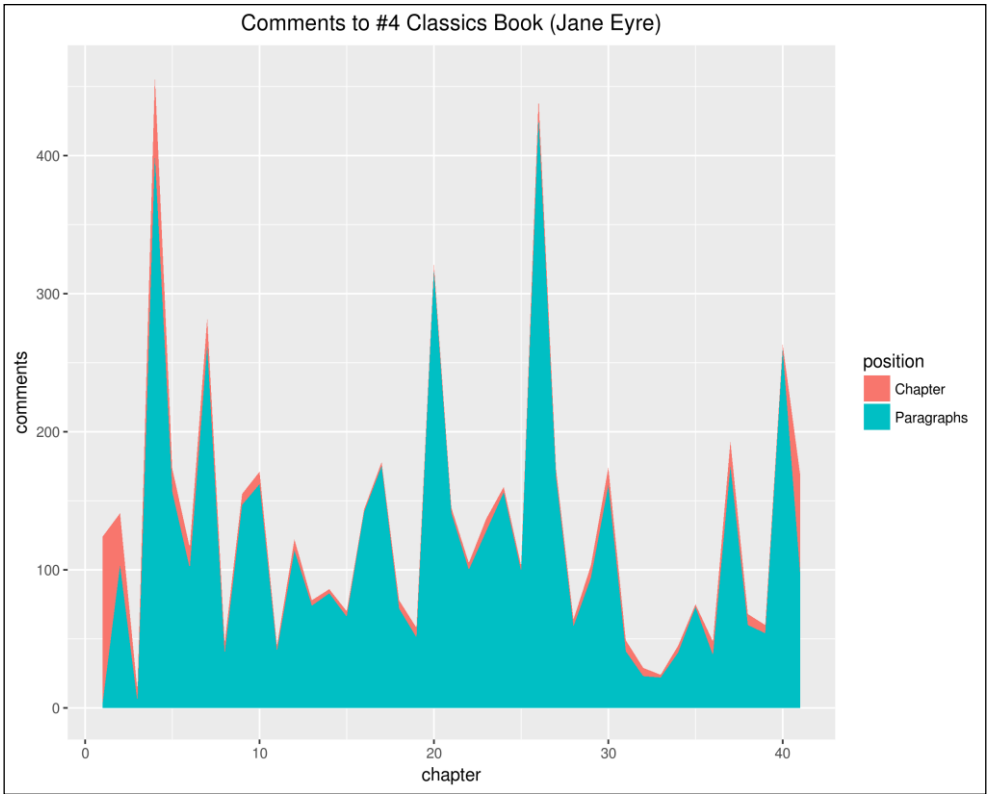


**Figure 5.** Number of Wattpad comments to *Jane Eyre*, divided per chapter.

## Wattpad as a New Resource for Sentiment Analysis

Sentiment Analysis (SA) is a technique that received a great impulse from marketing and socio-political applications (cf. Liu 2015) and has recently found an unexpected success in literary studies, too. Its popularity is mainly due to the parallel development of different researches and debates, like the one around the work of Matthew L. Jockers. After publishing a book that set the theoretical and methodological basis for a *Macroanalysis* of literary history (Jockers 2013), in May 2014 the co-founder of the *Stanford Literary Lab* published a post on his personal blog presenting a new method for the automatic detection of novels' plot (Jockers 2014). His idea was to use a SA algorithm that he developed (called *Syuzhet*, as a tribute to the narrative theories by Russian Formalists) to measure the positive/negative sentiment variation in narrative. Jockers used the *Portrait of the Artist as a Young Man* by James Joyce for his first test and showed how, in the high and low points of a graph, it was possible to identify the crucial plot twists of the novel. He later applied a clustering algorithm to a corpus of 41,383 novels, identifying six "archetypal plot shapes" for the emotional narrative arcs of the Western literary canon (Jockers 2015) and of contemporary blockbuster novels (Archer and Jockers 2016). These claims immediately raised some interest and disagreement, like for instance Annie Swafford's reply (2015), which thoroughly criticize Jockers's approach, by focusing on the *Syuzhet* algorithm's limitations and on the choice to group the narrative arcs using the Fourier transform (a mathematical process that "simplifies" graphs by eliminating the noise, but is also very sensitive to sudden trend variations). The controversy is not yet settled and Jockers's blog is continuously hosting new tests of the algorithm. One of the most illuminating remarks on the whole matter has been recently made by Adam Hammond:

> "Many distant reading projects have produced disappointing results because they have been more interested in validating their tools – showing that their computational methods are able to confirm existing stereotypes – than in pursuing genuine discoveries. Many others, meanwhile, produce provocative results that cannot be meaningfully validated" (Hammond 2017, 1)

Nevertheless, Sentiment Analysis is now spread to different areas and languages in the humanities (Sprugnoli et al. 2016; Zehe et al. 2017), different literary genres (Rebora 2017; Mellmann and Du 2018), and new research programmes like the neuro-aesthetics of literature (Jacobs et al. 2017). Moreover, in parallel to Jockers's research, other scholars came to similar results by using different instruments, identifying once again six "basic shapes" for the "emotional arcs of stories" (Reagan et al. 2016). And Kim, Padò and Klinger (2017) suggested to use a broader classification that is able to overcome the Manichaean positive/negative opposition in favour of a multidimensionality of emotions.

## The Experiment: Doubling Sentiment Analysis

In order to show the potential for SA of the Wattpad dataset, we designed a simplified experiment applying *Syuzhet* to the first and fourth most commented novels in the Classics category. The SA has been done on both the original text and the comments, and the resulting arcs have been compared. *Syuzhet* is not the most advanced SA software: for instance, the *SEANCE* algorithm (Crossley, Kyle, and McNamara 2017) offers an implementation of the multidimensionality suggested by Kim, Padò and Klinger (2017), while *Stanford SA* (Socher et al. 2013) mixes automated parsing of sentences and machine learning techniques (for a more extensive survey on these techniques, cf. Rojas-Barahona 2016). However, we chose *Syuzhet* for this preliminary test because of its popularity and for the advantage of implementing it in R, the same language used for the scraping. Its functioning is quite simple: based on a set of "emotional dictionaries" – which link words to sentiment values (e.g. "good" = +1; "bad" = -1) – the software counts the occurrences of these emotional tags in a chunk of text and returns a numeric output (e.g. "it was neither good nor bad" = 0). Such a simple mechanism leads to a range of inaccuracies, from the failure to detect irony and sarcasm to the impossibility to distinguish between affirmation and negation. Annie Swafford quotes the example "Well, it's like a potato", which is classified as extremely positive by *Syuzhet*, since words like "well" and "like" are erroneously interpreted as positive markers. Despite these limitations, *Syuzhet* is able to plot emotional narrative arcs that

 doi:10.4399/97888255181532

correspond quite accurately to the plot of the novel, especially if it is applied to broader chunks of text, rather than to single sentences.

The experiment has been implemented following Jockers's indications for the generation of a "simple_plot" with the "moving_average" function, which attributes a value to each chunk of text by calculating the average of all the surrounding chunks (Jockers 2017). This function balances the curve by normalizing all values within a range going from -1 to +1, otherwise it would be too noisy. The window size used to calculate the average is the default one, 10% of the novel's length[2], and we used the *Syuzhet* default sentiment dictionary. The most relevant changes that we introduced are:

1. instead of using the "get_sentences" function, we adopted the original division in paragraphs, in order to have a matching with the associated comments;

2. all the comments to a paragraph have been grouped in a single chunk, in order to maximize the number of words on which to calculate the sentiment value;

3. the sentiment values have been divided by the number of words on which they have been calculated, in order to normalize them on the same scale;

4. the steps on the x axis have been set based on paragraph length, thus longer paragraphs are represented by a wider step.

The analyses of the novels *Pride and Prejudice* and *Jane Eyre*, with the respective comments, are visualised in Figure 6 and Figure 7.

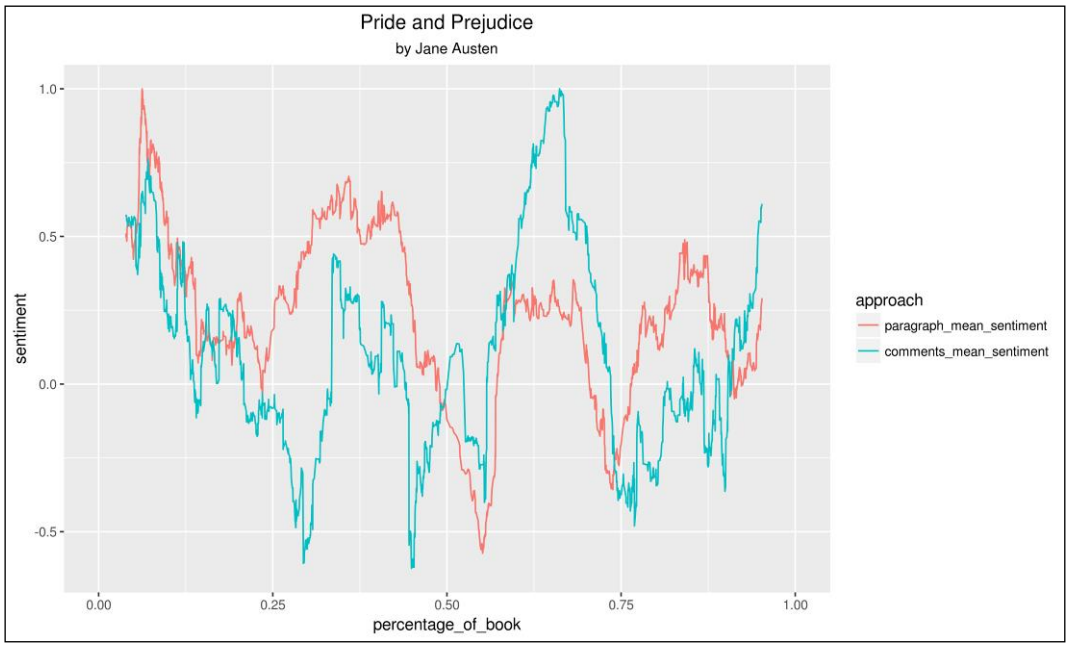


**Figure 6.** Emotional arcs of *Pride and Prejudice*, based on paragraphs and comments.

[2] A consequence of this choice is the exclusion of the first and last 5% of the novel from the graphs.

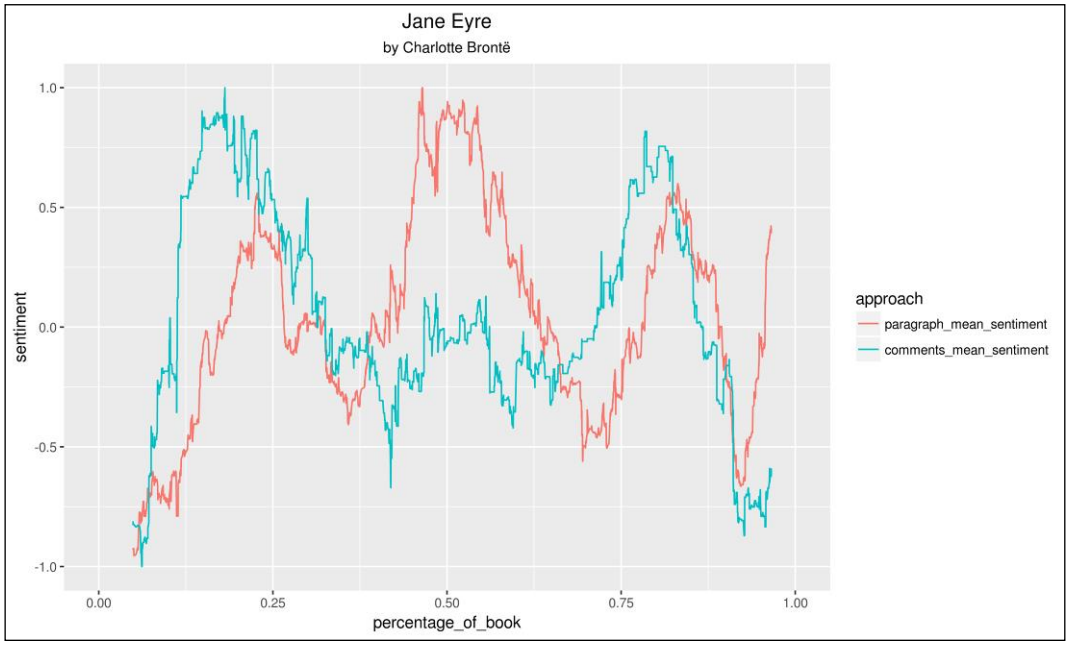


**Figure 7.** Emotional arcs of *Jane Eyre*, based on paragraphs and comments.

## Discussion

The emotional arcs calculated on the original text and on the comments show similar general trends, but also strong discrepancies. This is due to the limitations of the algorithm but also to the diversity of the two datasets: on the one hand, the SA calculated on thousands of words (for the comments) is more reliable than an analysis done on a few dozen words (for the paragraphs); on the other hand, comments are not necessarily an explanation or paraphrase of the text – although sometimes this is the case – and they can also have a social function not directly related to the novel. Anyway, at a later stage, it will be interesting to see what paragraphs elicited the most positive or negative responses, and to determine whether these responses have been triggered by a plot event, by some language use, or by the discussion with other readers.

A significant example of the limitations/potentialities of SA is offered by the portion of text surrounding the 30.7 "percentage_of_book" of *Pride and Prejudice* (cf. Figure 6)[3], that is the point where the discrepancy between the sentiment of paragraph and comments reaches its maximum. *Syuzhet* attributes a value of +0.57 to the paragraph and of -0.49 to the comments. The first value would seem more appropriate, considering that the segment corresponds to Chapter XXI – following Elizabeth's refusal to marry the insensible Mr. Collins – but a deeper investigation reveals that the situation is more complex. Figure 8 shows the words that determined the sentiment values: the differences between the two sections of the wordcloud (note the decreased relevance of a concept like "happiness" in the comments, where many new semantic areas also appear) confirm how comments cannot be simply used for a better-refined SA of the text, because they depict a quite different phenomenon.

[3] In order to emulate the "moving_average" procedure, we selected the surrounding 10%.


**Figure 8.** Comparative wordcloud of the words tagged by *Syuzhet* as positive and negative in paragraphs and comments.

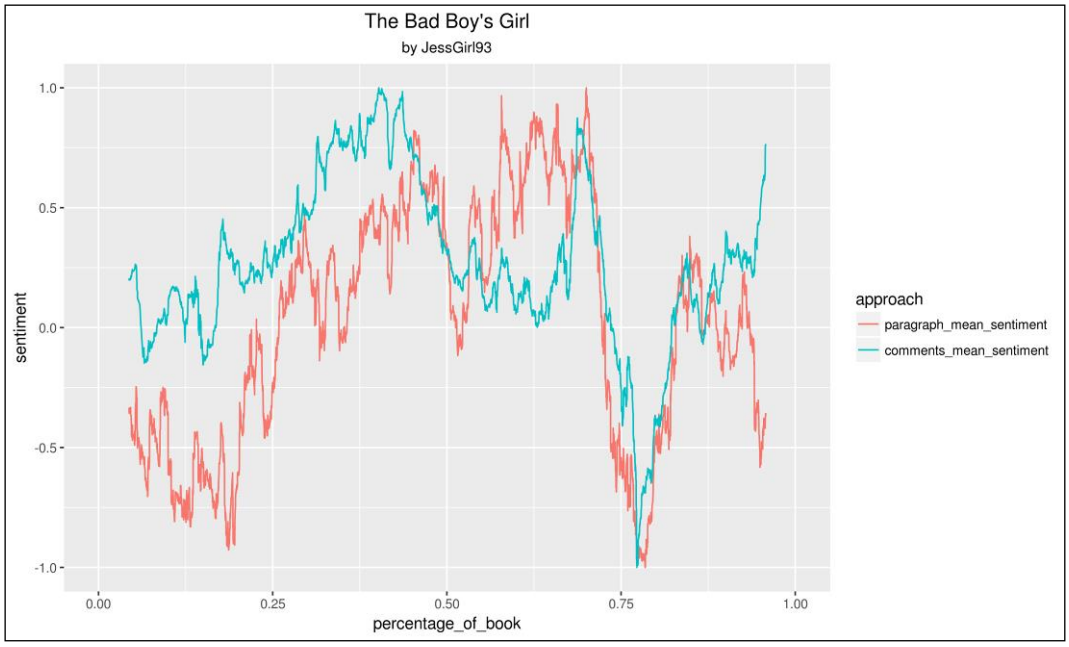


**Figure 9.** Emotional arcs of *The Bad Boy's Girl*, based on paragraphs and comments (we excluded the last chapters, which advertise the next book in the series).

In addition, it is worth underlining that the generation of emotional arcs is an operation that does not have a unique output. The research group led by Andrew Reagan is providing a free online tool (*Hedonometer*) for the generation of interactive graphs for famous novels in the Western literary canon. The algorithm and the sentiment dictionary are different, and so are the results[4]. In our case, also, it is probable that the potential provided by the large amount of comments is only partially exploited for the Classics section: for example, it is striking that the central part of Jane Eyre (corresponding to Jane's engagement with Mr. Rochester) does not trigger a positive sentiment in the comments. This is mainly due to the fact that the comments to that section are just a few, not enough to change the trend in the graph – even if we cannot overlook that Wattpad

[4] See for example the *Pride and Prejudice* graph and the *Jane Eyre* graph.

readers are used to more intense and explicit depictions of passionate love (see the books in the Teen Fiction category). We chose to analyse *Jane Eyre* because of its even distribution of comments along the novel, but the 160 comments to chapter XXI provide less data than its 195 paragraphs. In the case of very commented texts, like the first Teen Fiction title (Figure 9), the correspondences between the emotional arcs of the story and of the comments are much closer (see how close the values for the most negative lows are). However, there are differences also in this case, in particular for the incipit and the conclusions, therefore we cannot neglect that the emotions represented in the text will always be somehow incommensurable to the readers' emotions.

We acknowledge that SA has many intrinsic limitations, but we also showed the great potential it can have, especially when we can rely on numbers getting closer to the scale of "big data." This is just the very first step of a promising and exciting research programme. We tried to suggest some possible developments that we hope will set a new and so far unexplored perspective to approach some crucial topics of literary theory and criticism.

## References

Archer, Jodie, and Matthew L. Jockers. *The Bestseller Code : Anatomy of the Blockbuster Novel*. New York: St. Martin's Press, 2016.

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55.4 (2012). ACM: 77–84. doi:10.1145/2133806.2133826.

Bleich, David. *Subjective Criticism*. Baltimore: The Johns Hopkins University Press, 1978.

Caracciolo, Marco. *The Experientiality of Narrative: An Enactivist Approach*. Berlin: de Gruyter, 2014.

Cordón-García, José-Antonio, Julio Alonso-Arévalo, Raquel Gómez-Díaz, and Daniel Linder. *Social Reading*. Oxford: Chandos, 2013.

Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. "Sentiment Analysis and Social Cognition Engine (SEANCE): An Automatic Tool for Sentiment, Social Cognition, and Social-Order Analysis." *Behavior Research Methods* 49.3 (2017): 803–21. doi:10.3758/s13428-016-0743-z.

Davies, Rosamund. "Collaborative Production and the Transformation of Publishing: The Case of Wattpad." In *Collaborative Production in the Creative Industries*, edited by James Graham and Alessandro Gandini, 51–67. Westminster: University of Westminster Press, 2017.

Dilevko, Juris, and Keren Dali. "The Self-Publishing Phenomenon and Libraries." *Library & Information Science Research* 28.2 (2006). JAI: 208–34. doi:10.1016/J.LISR.2006.03.003.

Fast, Ethan, Tina Vachovsky, and Michael S. Bernstein. "Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community." In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, 112–20. Available at http://arxiv.org/abs/1603.08832

Fellbaum, Christiane. "WordNet." In *Theory and Applications of Ontology: Computer Applications*, 231–43. Dordrecht: Springer Netherlands, 2010. doi:10.1007/978-90-481-8847-5_10.

Graham, James, and Alessandro Gandini. *Collaborative Production in the Creative Industries*. Westminster: University of Westminster Press, 2017. doi:10.16997/book4.

Hammond, Adam. “The Double Bind of Validation: Distant Reading and the Digital Humanities’ ‘Trough of Disillusionment.’” *Literature Compass* 14.8 (2017): 1–13. doi:10.1111/lic3.12402.

Herrmann, J. Berenike. “In a Test Bed with Kafka. Introducing a Mixed-Method Approach to Digital Stylistics.” Digital Humanities Quarterly 11.4 (2017). Available at: http://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html.

Hoffelder, Nate. “Damn the Facts: The 'Ebook Sales Are Down' Narrative Must Be Maintained at All Costs.” *The Digital Reader* (2017). Available at http://the-digital-reader.com/2017/04/27/damn-facts-ebook-sales-narrative-must-maintained-costs/.

Iser, Wolfgang. *The Act of Reading : A Theory of Aesthetic Response*. London: Routledge & Kegan Paul, 1978.

Jacobs, Arthur M., Sarah Schuster, Shuwei Xue, and Jana Lüdtke. “What’s in the Brain That Ink May Character ….” *Scientific Study of Literature* 7.1 (2017): 4–51. doi:10.1075/ssol.7.1.02jac.

Jockers, Matthew L. *Macroanalysis : Digital Methods and Literary History*. Urbana: University of Illinois Press, 2013. Available at: https://books.google.it/books/about/Macroanalysis.html?id=mPOdxQgpOSUC&redir_esc=y.

Jockers, Matthew L. “A Novel Method for Detecting Plot.” (2014). Available at: http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/.

Jockers, Matthew L. “The Rest of the Story.” (2015). Available at: http://www.matthewjockers.net/2015/02/25/the-rest-of-the-story/.

Jockers, Matthew L. “Introduction to the Syuzhet Package.” *The Comprehensive R Archive Network* (2017). Available at: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html.

Kim, E, S. Padó, and R. Klinger. “Investigating the Relationship between Literary Genres and Emotional Plot Development.” In *Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings*, 17–26. Association for Computational Linguistics, 2017. Available at: http://www.aclweb.org/anthology/W/W17/W17-22.pdf#page=31.

Kukkonen, Karin, and Marco Caracciolo. “Introduction: What Is the ‘Second Generation?’” *Style* 48.3 (2014): 261–74. doi:10.5325/style.48.3.261.

Li, Wu, Xingxing Wu, and Anhui Wang. “To Stick or to Switch: Understanding Social Reading Apps Continuance by Evidence Collected from Chinese College Students.” In *New Media and Chinese Society*, edited by Ke Xue and Mingyang Yu, 223–37. Singapore: Springer, 2017. doi:10.1007/978-981-10-6710-5_13.

Liu, Bing. *Sentiment Analysis*. Cambridge: Cambridge University Press, 2015. doi:10.1017/CBO9781139084789.

Mellmann, Katja, and Keli Du. “Sentimentanalyse in Unstrukturierten Texten (Am Bsp. Literaturgeschichtlicher Rezeptionsanalyse).” In *DHd 2018 Konferenzabstracts*, 305–8. Cologne: Universität zu Köln, 2018. Available at: http://dhd2018.uni-koeln.de/programm-freitag/.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. “Distributed Representations of Words and Phrases and Their Compositionality.” *Proceedings of the*

*26th International Conference on Neural Information Processing Systems - Volume 2*. Curran Associates Inc., 2013. Available at: https://dl.acm.org/citation.cfm?id=2999959.

Miller, Monica. “What Wattpad Brings to the Publishing Table." *PUB800* (Fall 2015). Available at: https://tkbr.publishing.sfu.ca/pub800/2015/12/what-wattpad-brings-to-the-table/.

Mirmohamadi, Kylie. *The Digital Afterlives of Jane Austen: Janeites at the Keyboard*. Basingstoke: Palgrave Macmillan, 2014. doi:10.1057/9781137401335.0001.

Nakamura, Lisa. “‘Words with Friends’: Socially Networked Reading on Goodreads.” *Pmla* 128.1 (2013): 238–43. doi:10.1632/pmla.2013.128.1.238.

Passalacqua, Franco, and Federico Pianzola. “Epistemological Problems in Narrative Theory: Objectivist vs. Constructivist Paradigm.” In *Narrative Sequence in Contemporary Narratology*, edited by Raphaël Baroni and Françoise Revaz, 195–217. Columbus: Ohio State University Press, 2016.

Pianzola, Federico. “Cognitive Affordances, Aesthetic Effects and Social Functions: A Systemic Approach to Narrative Studies.” *Culture, Biography & Lifelong Learning* 3.3 (2017).

Ramdarshan Bold, M. “The Return of the Social Author: Negotiating Authority and Influence on Wattpad.” *Convergence: The International Journal of Research into New Media Technologies* (2016). doi:10.1177/1354856516654459.

Ramdarshan Bold, Melanie, and Kiri L. Wagstaff. “Marginalia in the Digital Age: Are Digital Reading Devices Meeting the Needs of Today’s Readers?” *Library & Information Science Research* 39.1 (2017): 16–22. doi:10.1016/J.LISR.2017.01.004.

Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. “The Emotional Arcs of Stories Are Dominated by Six Basic Shapes.” *EPJ Data Sci.* 5 (2016): 5–31. doi:10.1140/epjds/s13688-016-0093-1.

Rebora, Simone. “A Software Pipeline for the Reception of Italian Literature in Nineteenth-Century England.” In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017*, 129–34. New York, New York, USA: ACM Press, 2017. doi:10.1145/3078081.3078102.

Rojas-Barahona, Lina Maria. “Deep Learning for Sentiment Analysis.” *Language and Linguistics Compass* 10.12 (2016): 701–19. doi:10.1111/lnc3.12228.

Rosenblatt, Louise M. *The Reader, the Text, the Poem: The Transactional Theory of the Literary Work*. Carbondale: Southern Illinois University Press, 1978. Available at: https://muse.jhu.edu/book/42573.

Rowberry, S. P. “Commonplacing the Public Domain: Reading the Classics Socially on the Kindle.” *Language and Literature* 25.3 (2016): 211–25. doi:10.1177/0963947016652782.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.” *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013): 1631–42. Available at: https://aclanthology.coli.uni-saarland.de/papers/D13-1170/d13-1170.

Sprugnoli, Rachele, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. “Towards Sentiment Analysis for Historical Texts.” *Digital Scholarship in the Humanities* 31.4 (2016): 762–72. doi:10.1093/llc/fqv027.

Stein, Bob. “A Taxonomy of Social Reading: A Proposal.” (2010). Available at: http://futureofthebook.org/social-reading/.

Sternberg, Meir. “Telling in Time ( II ): Chronology , Teleology , Narrativity.” *Poetics Today* 13.3 (1992): 463–541.

Swafford, Annie. “Problems with the Syuzhet Package.” *Anglophile in Academia: Annie Swafford’s Blog* (2015). Available at: https://annieswafford.wordpress.com/2015/03/02/syuzhet/.

Thelwall, Mike, and Kayvan Kousha. “Goodreads: A Social Network Site for Book Readers.” *Journal of the Association for Information Science and Technology* 68.4 (2017): 972–83. doi:10.1002/asi.23733.

van Peer, Willie, Frank Hakemulder, and Sonia Zyngier. *Scientific Methods for the Humanities*. *Linguistic Approaches to Literature*. Amsterdam: John Benjamins, 2012. doi:10.1075/lal.13.

Vlieghe, Joachim, Jaël Muls, and Kris Rutten. 2016. “Everybody Reads: Reader Engagement with Literature in Social Media Environments.” *Poetics* 54: 25–37. doi:10.1016/j.poetic.2015.09.001.

Weinrich, Harald. *Tempus : Besprochene Und Erzählte Welt*. München: Beck, 2001.

Williams, Abigail. *Social Life of Books : Reading Together in the Eighteenth- Century Home*. New Haven & London: Yale University Press, 2017.

Winders, James A. *Gender, Theory, and the Canon*. Madison: The University of Wisconsin Press, 1991. Available at: https://uwpress.wisc.edu/books/0286.htm.

Zehe, Albin, Martin Becker, Fotis Jannidis, and Andreas Hotho. “Towards Sentiment Analysis on German Literature.” In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, 387–94. Cham: Springer, 2017. doi:10.1007/978-3-319-67190-1_36.