Blurb Mining : Discovering Interesting Excerpts from E-commerce Product Reviews

Saratchandra Indrakanti
eBay Inc.
San Jose, California
sindrakanti@ebay.com

Gyanit Singh
eBay Inc.
San Jose, California
gysingh@ebay.com

Justin House
eBay Inc.
San Jose, California
juhouse@ebay.com

ABSTRACT

Product reviews on modern e-commerce websites have evolved into repositories of valuable firsthand opinions on products. Showcasing the opinions that reviewers express on a product in a succinct way can not only promote the product, but also provide an engaging experience that simplifies the shopping journey for online shoppers. In the case of traditional media such as movies and books, employing *blurbs* or excerpts from critic reviews for promotional purposes is an established practice among movie publicists and book editors that has proven to be an effective way of capturing attention of customers. Such excerpts can be discovered from e-commerce product reviews to highlight interesting reviewer opinions and add emotive elements to otherwise bland e-commerce product pages. While traditional movie or book blurbs are manually extracted, they must be automatically extracted from e-commerce product reviews owing to the scale of catalogues. Further, traditional blurbs are generally phrased to be very positive in tone and sometimes may take some words out of context. However, excerpts for e-commerce products must represent the true opinions of the reviewers and must capture the context in which the words were used to retain trust of users. To that end, we introduce the problem of extracting engaging excerpts from e-commerce product reviews in this paper. We present methods to automatically discover such excerpts from reviews at scale by leveraging natural language properties such as syntactic structure of sentences and sentiment, and discuss some of the underlying challenges. We further present an evaluation of the effectiveness of the proposed methods in terms of the quality of the blurbs generated and their ranking orders produced.

CCS CONCEPTS

• **Information systems** → **Information extraction**; *Sentiment analysis*; *Summarization*; • **Computing methodologies** → *Language resources*;

KEYWORDS

Blurbs, Opinion Mining, Product Reviews, Aspect Mining, E-commerce, Sentiment Analysis

ACM Reference Format:

Saratchandra Indrakanti, Gyanit Singh, and Justin House. 2018. Blurb Mining : Discovering Interesting Excerpts from E-commerce Product Reviews . In

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '18 Companion, April 23-27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5640-4/18/04.

<https://doi.org/10.1145/3184558.3191626>

WWW '18 Companion: The 2018 Web Conference Companion, April 23-27, 2018, Lyon, France. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3184558.3191626>

1 INTRODUCTION

Employing excerpts from critic reviews or *blurbs* as a promotional tool has been a long-established practice in the publishing and media industries. Movie watchers and book readers often come across excerpts such as “*The must-see movie of the season*” in movie advertisements or “*Every single word has immense value*” on book covers, attributed to seasoned critics in the field whom the target audience trust. Movie publicists and book editors meticulously select such excerpts that speak about certain characteristics of the subject that appeal to potential audience and pique their interest [5]. In recent years, websites of various consumer products have been quoting excerpts from expert reviews as promotional verbiage for the products. The combination of interesting excerpts, or blurbs, that entice the audience along with the credibility of the review author, can capture customer attention to a product.

The concept of blurbs, which have traditionally been used in media and publishing industries, easily lends itself to modern e-commerce websites. Modern e-commerce websites have vast catalogues of products listed on their platforms, offering their customers with a wide array of choices. Many of these products attract a large number of reviews from buyers who share their firsthand experiences in using the product. Online shoppers have increasingly been relying on product reviews to make their buying decisions. Nevertheless, owing to the large selection of products listed on these websites and the growing content on the product pages, it becomes increasingly unrealistic to expect shoppers to read all the reviews for a product. This necessitates introducing techniques to showcase the opinions about a product in a succinct and personable way to not only capture user attention, but also to improve their shopping experience. To that end, blurbs which have been extensively used in the media industry, are very well-suited to e-commerce products. For instance, “*Great story and magnificent graphics!*”, a blurb from reviews for a video game, or “*Images are super sharp, great noise control*”, a camera blurb, provide concise yet informative glimpses of the products.

Conventionally, blurbs are meticulously hand-picked by publicists from critic reviews. However, this is not possible in the case of e-commerce products, owing to the scale of modern e-commerce platforms which may have hundreds of millions of products listed in their catalogues. Methods must be developed to automatically discover and extract blurbs from e-commerce product reviews at scale. There are several distinctions that blurb mining methods must pay attention to. First, while movie or book blurbs are known

Figure 1: Illustration of a blurb showcased on an e-commerce product page

to paraphrase the original reviews [2], and occasionally represent words out of context, it is critical that e-commerce product blurbs ensure that the context of an excerpt is retained and refrain from exaggeration in order to preserve user trust in the platform. Further, in the case of an e-commerce website, the goal is to find enticing and representative blurbs for a given product, rather than promoting one product over another. In addition to this, blurbs don't necessarily have to be entire sentences, but can be succinct excerpts of sentences ensuring they retain the context, unlike many existing extractive summarization tasks. Specifically, taking note of the observation that reviewers often write long sentences where they may discuss multiple topics, it is important to identify excerpts of sentences to ensure that blurbs are succinct and focused on a topic. In addition to this, blurbs must comprise of product-specific opinions rather than generic statements, so that they can be informative to online shoppers. For instance, the excerpt *"great lens, love the low light ability"* is more informative than a generic excerpt such as *"I love this product a lot"*

In this paper, we introduce the problem of automatically discovering blurbs from product reviews. We will describe the methodology developed and discuss the challenges involved in achieving this at scale. To mine blurbs from product reviews, we identify excerpts containing product-specific opinions and subsequently rank them to discover the best ones. We employ a three-phase process that includes aspect mining, excerpt extraction and excerpt ranking to achieve this. We first discover product aspects, or attributes of a product discussed in the reviews, as part of aspect mining, by aggregating word dependencies from sentence dependency structure of review sentences. We retain those sentences where opinions have been expressed about the identified aspects. Next, we extract excerpts from those sentences while retaining the context as part of excerpt extraction. A supervised learning model is then trained to rank the excerpts based on wisdom of crowd, popularity of topics and sentence quality to pick most accurate, clear and agreed upon excerpts. We present and evaluate the effectiveness of the proposed methods in terms of the quality of the extracted blurbs and the ranking generated, in the evaluations section.

The problem of blurb mining for e-commerce products may have some commonalities with extractive summarization and opinion mining, but also differs in several ways, necessitating specialized methods to be developed. Extractive summarization involves identifying the most important sentences in a body of text, by scoring sentences based on an application-specific measure of importance. While blurb mining involves ranking excerpts, it differs in a few main ways. First, in extractive summarization, sentences are scored,

whereas blurbs are not necessarily complete sentences, but excerpts of sentences. Next, blurbs must capture product-specific opinions and may not be useful as generic sentences. Further, unlike extractive summarization, the goal of blurb mining is not generating a summary of reviews, but discovering the most informative, enticing and succinct excerpts. In other words, blurb mining is aimed at narrowness but succinctness of excerpts, rather than completeness of topics in a summary. Opinion mining focuses on extracting opinions and opinion-targets from text. Similar to opinion mining, blurb mining focuses on identifying product-specific opinions and product aspects. However, in addition to that, blurb mining involves extracting excerpts from sentences that contain the opinions and product aspects. Further, the extracted opinionated excerpts must be ranked to discover the most impactful blurbs that convey popular and agreed upon topics. Hence, we propose methods specifically focused on discovering interesting blurbs from e-commerce product reviews.

The contributions of this paper can be summarized as follows:

- We introduce the novel problem of blurb mining from e-commerce product reviews, and discuss the challenges involved in developing a scalable solution and its potential impact.
- We propose scalable methods to discover blurbs from product reviews that are enticing, product-specific and succinct.

2 RELATED WORK

Blurbs have been extensively used in publicizing traditional media such as books and movies. The authors of [7] and [8] have provided some overview of the usage of blurbs in the book publishing industry. The idea of blurbs has been adopted by movie publicists and more recently promoters of consumer products on the product webpages, by quoting expert or critic reviews. However, the blurbs in this context are manually selected and phrased by publicists or editors. We are not aware of any automated processes of blurb generation at scale or methodologies published in that regard.

Nevertheless, online customer reviews including movie reviews, restaurant reviews and e-commerce product reviews have received considerable interest from the research community. Product reviews on e-commerce websites, which are the subject of this article, have been extensively studied in general and by the text mining community specifically. Several researchers have focused on opinion mining [12], sentiment analysis [24] [15] and summarization of product reviews [1] [30]. Specifically, product feature based opinion mining and extractive summarization are areas which share commonalities with blurb mining.

Product review blurbs are expected to discuss product-specific aspects and hence build upon product feature based opinion mining. In the field of opinion mining, several product aspect extraction models that are based on reviewer opinions have been presented by researchers over the past decade. Various supervised, unsupervised and semi-supervised models have been proposed to identify opinionated product aspects from reviews. Authors of [10] first introduced an unsupervised model based on frequent itemset mining. Other unsupervised methods include leveraging PMI statistics [10], a rule based approach that employs linguistic patterns on sentence dependency trees [22], a graph based method that captures word

dependencies[29], and a model to automatically infer rules to be applied on dependency trees [13], among several others. Different variations of topic models including multi-grain topic models that extends topic modeling techniques such as LDA and PLSI to extract local and global aspects for products [25], and probabilistic graphical models that extend PLSI [17]. Semi-supervised models such as [23], [18], [28] have been introduced to guide certain unsupervised models towards more precise aspects by using a domain specific set of seed words. Several supervised models have been proposed to extract aspects from reviews, such as [11] based on conditional random fields and a convolutional neural network based model [21]. The authors of [27] presented a method based on aspect ranking and grouping to assist sellers in phrasing promotional statements for products on e-commerce websites. We adopt an unsupervised approach to discovering aspects that can be used for blurb generation, in order to develop a method that scale to a large number of diverse e-commerce product categories.

Extractive summarization is another area of text mining that is related to some of the tasks involved in blurb mining. Extractive summarization techniques have been developed for a variety of domains including online review sentiment summarization [26], restaurant domain [20], biomedical domain [16] and student course feedback [14]. A survey of extractive text summarization techniques is provided in [19]. One of the main tasks in the area of extractive summarization is identifying the most important sentences in a body of text by an appropriate measure of importance. A comparative study of several sentence scoring techniques is presented in [6].

3 METHODOLOGY

In this section, we describe in detail the 3-phase approach developed to mine blurbs from e-commerce product reviews. Specifically, we describe the aspect mining, excerpt extraction and excerpt ranking steps. We first discover product aspects, or vocabulary upon which product-specific opinions have been expressed using the methodology described in aspect mining. We then retain those review sentence that have occurrences of the discovered product aspects. Review sentences often happen to be long, and contain several different topics discussed and opinions expressed. To identify succinct and topic-focused portions of sentences, we apply excerpt extraction techniques that return excerpts of sentences that capture the context of the corresponding product aspect and opinion. Supervised learning models trained to discover interesting excerpts are applied to rank the excerpts identified and discover interesting blurbs. An overview of this methodology is provided in Figure 2.

3.1 Aspect Mining

Excerpts that capture product-specific opinions are more informative to users than generic statements. To identify such product-specific excerpts we first discover product aspects from review content. Aspects are attributes or features of a product discussed in reviews, upon which the reviewer expresses an opinion. They are also referred to as opinion targets in the field of opinion mining. For instance, *shutter speed*, *aperture*, *zoom range*, could be aspects discussed in camera reviews. Consider the following excerpt from a review for a video game: “*Compelling gameplay and story and*

beautiful graphics”. The reviewer expresses the opinion *compelling* on the aspect *gameplay*. The challenge in discovering aspect candidates mainly involves identifying those words that 1) the reviewer has expressed an opinion on, and 2) are attributes of the product and describe its features. Further, to quantify the importance of the identified aspects, the collective opinions expressed by reviewers on the aspects must be aggregated. Thus, this task entails firstly, identifying opinion and target word candidates, and next, aggregating them to quantify reviewer emphasis.

We exploit the properties of sentence dependency trees to extract aspect candidates from each sentence of every review. Sentence dependency trees capture the grammatical relations between words that comprise a sentence. The dependencies are binary asymmetric relations between a word identified as head (generally a verb) and its dependents [4]. The nature of the relationship is denoted by a dependency label associated with the edge connecting the two words in the relationship. For instance, Figure 3 depicts the dependency tree for the following sentence: “*Camera is average but it will take pretty good photos at night.*” A detailed description of dependency trees can be found in [4]. The open source library Spacy [9] is employed to generate dependency trees owing to its combination of speed and accuracy [3]. Punctuation is retained and no pre-processing is performed prior to producing dependency trees, since lemmatization or other pre-processing steps may affect the accuracy of dependency tree generation.

From a dependency tree constructed for a given sentence, we select those dependencies that capture the aspect candidates. Specifically, we identify the dependencies that capture the relations between the aspect and the opinion words associated with the aspect. Aspect candidates are identified subject to a set of dependency rules R described below, some of which have been defined by the authors of [22], that determine if a word is an aspect candidate.

- (1) A noun n is an aspect if there exists a parent-child relation between n and another word a , and a is either an adjective or an adverb. Here the set of words a that satisfy this condition constitute the opinion.
- (2) Adjective-adverb sibling rule: A noun n is selected as an aspect, and a as the opinion if there exists a word a that shares the same parent (head term), and a is either an adjective or adverb.
- (3) If a word v has a direct object relation with a noun n , and v is a verb, then n is selected as an aspect, with v being the opinion.
- (4) If A verb v has a noun n as a parent (head term), and v has a parent-child relation with another word a which is an adjective, then n is selected as an aspect, with a being the opinion.
- (5) If a noun n_1 in a conjunct relation or a prepositional relation with another noun n_2 , and n_2 is parent-child relation with an adjective a , n_1 is selected as an aspect with a being
- (6) If a word v is in an open clausal relationship with another word a , with v being a verb and a being an adjective or adverb, then select v as an aspect and a as opinion.

Given a sentence s , its dependency tree $D_s : \{dep\langle w_1, w_2 \rangle\}$, where $w_1, w_2 \in s$ are any two words in the sentence s with a dependency relation dep , is generated. Aspect candidates α satisfying

Figure 2: Overview of methodology for blurb discovery

Figure 3: Sentence dependency tree generated for an example review sentence “Camera is average but it will take pretty good photos at night.”. Nodes are individual words in a sentence, while edges represent dependencies between words.

atleast one of the rules in R are selected along with the associated opinion words and dependencies. The satisfying relations $dep \langle O_\alpha, \alpha \rangle$ are returned, where O_α is the opinion term associated with α . Figure 4 shows the selected dependencies for the example sentence. The selected aspect candidates and their associated opinions are aggregated to quantify their importance as collectively expressed by the reviewers. Frequency of occurrence of aspect candidates is a simple measure of their importance. Nevertheless, this can be extended to measures such as aggregate sentiment associated with an aspect candidate. A set of aspects, along with a measure of their importance is generated.

Figure 4: Dependency relations returned after applying the dependency tree pruning algorithm on the example sentence.

3.2 Excerpt Extraction

As mentioned previously, blurbs must contain product-specific content, while being short in length. To that end, review sentences that have occurrences of aspects are retained to be considered for blurb generation. However, many of those sentences may be long, necessitating methods to extract excerpts that discuss the aspect and the associated opinion, yet, are coherent to read. In this section, we describe an excerpt extraction technique that identifies coherent excerpts from a sentence that are not longer than a specified length and capture a given aspect and associated opinion.

The goal of constituency parsing (also known as “phrase structure parsing”) is to identify the phrases in the text. Constituent structure is based on the observation that words combine with other words to form units. Constituency parse trees organize syntax into constituents.

Figure 5: Illustration of excerpt extraction based on sentence parse tree. Excerpt, circled in red, consisting of aspects *weight and balance* is extracted from the parse tree for sentence “The weight and balance of the club are good for my swing and it gives me a sense of confidence”

The proposed excerpt extraction technique utilizes the constituency sentence parse tree of a sentence to identify coherent excerpts. Constituency parse trees organize the syntactical structure into constituents based on the observation that words combine with other words to form linguistic structures. They assist in identifying phrases in text. Stanford parser is used to generate the constituency parse tree for a given sentence. Figure 5 shows a parse tree generated for the sentence “The weight and balance of the club are good for my swing and it gives me a sense of confidence”. Given an aspect α and the associated opinion O_α that has been expressed in the sentence, and a specified upper bound k on the number of words permissible in an excerpt, the subtree τ with the maximum number of leaves that confirms to the following conditions is returned.

- (1) The root of the subtree τ must be either a sentence or clause (S) or a noun-phrase (NP)
- (2) The subtree τ may have at the most k leaves
- (3) Aspect α and opinion O_α are leaves of the subtree τ

To identify the subtree to be returned, first the aspect and opinion leaves are retrieved from the tree and their common ancestors are discovered such that the above rules are satisfied. The subtree rooted at the selected common ancestor is returned, and its leaves constitute the extracted excerpt. For the example sentence depicted in figure 5, *weight* and *balance* are aspects on which the opinion *good* has been expressed, and let $k = 15$ for the purpose of illustration. The subtree to be returned must have at the most 15 leaves, root of the subtree must either be S or NP , and must contain the leaves *weight*, *balance* and *good*. The subtree highlighted in red in figure 5 is returned which represents the excerpt “*The weight and balance of the club are good for my swing*” for this sentence.

3.3 Excerpt Ranking

We intend to discover the excerpts that could be the most interesting to users, among the excerpts candidates, to be presented as blurbs. We train a supervised learning model to assist in ranking the excerpts by leveraging crowd wisdom to represent popularity of topics and sentence quality to pick most interesting, accurate and agreed upon excerpts. The model is trained to identify excerpts that are interesting against those that aren’t. The learning problem is formulated as a classification problem, where the learner is trained to classify an excerpt as *interesting*, *neutral* or *not interesting*. Several syntactic, grammatical, sentiment-based and review-specific features are used in training the model. Parts of speech tags of words in the original review sentence and excerpt are used to build features that capture the numbers of adjectives, adverbs, nouns and verbs. Features based on dependency relations between aspect and opinion terms in the excerpt provide syntactic information. Text features such as character length, number of words in the sentence and the excerpt, tf-idf scores of the corresponding aspect and its frequency of occurrence among excerpt candidates are adopted. Further, the positive, negative and compound sentiment of the sentence are used to capture the reviewer sentiment. In addition to this, product and review-specific features such as rating provided by the reviewer, overall rating for the product and numbers of positive and negative ratings for the product are used. Specifically, features related to aspect-pertinent metrics within the reviews, and those that quantify sentiment are intended to assist the classifier in learning excerpts that are interesting. A gradient boosted tree based classifier is trained to classify the excerpts to the afore-mentioned classes. The classifier prediction for an excerpt to belong to the *interesting* class is used as a measure for ranking the excerpts.

4 EVALUATION

In this section, we present evaluation of the proposed methods for blurb extraction and ranking. Specifically, we evaluate the quality of the extracted blurbs and effectiveness of the ranking determined for those blurbs. To generate a ranked set of blurbs for each product, first, product-specific aspects are discovered from the reviews for that product. Next, excerpts with a specified upper bound on length (8 words in the evaluations that will be presented in this section), are

extracted from review sentences containing the discovered aspects. Those excerpts are ranked with the assistance of a classifier trained on a human annotated e-commerce blurb corpus. This process is illustrated along with some examples of blurbs generated in Table 1.

To facilitate the evaluation, we curated an evaluation dataset consisting of blurbs generated for 25 products sampled from 20 different e-commerce product categories including those such as laptops, smartphones, books, movies and sporting goods. The blurbs for each of these products were labeled by 3 independent human evaluators. The evaluators were asked to label the blurbs as either *interesting*, *neutral* or *not interesting*, consistent with the labels used in the excerpt ranking phase. Blurbs that are irrelevant, improperly constructed or lacking proper context were labeled as *not interesting*. The evaluators were provided with the product title and description. The majority consensus was adopted as the label for a blurb.

First, we evaluate the quality of the extracted blurbs. We determine the proportion of blurbs that have been classified as *interesting*, that the evaluators labeled as *interesting*. Precision@ k values are computed for the evaluation dataset, where k is the number of top blurbs considered in the calculation. Figure 6(a) shows the precision@ k values obtained for $k = [2, 14]$ ranging from 0.71 to 0.48. Next, we evaluate the quality of the ranking order generated for the blurbs. We use the NDCG metric for this purpose. Figure 6(b) shows the NDCG@ k values obtained for the evaluation dataset with $k = [2, 14]$. The NDCG values range from 0.94 to 0.88, as can be observed from the plot.

The precision and NDCG values obtained indicate that, while the ranking order produced for the blurbs discovered is satisfactory, there is room for improvement in the quality of the blurbs extracted. An analysis of the false positives that occurred in blurb extraction reveals two main sources of false positives. First, excerpts that are not very product-specific but are generic in nature may not be informative to online shoppers are not considered to be interesting blurbs. For instance, the excerpt “*I highly recommend this computer*” from a product review for a laptop does not provide an opinion on specific features of the laptop. Such false positives are a consequence of false positives in aspect mining. The second category of false positives are those that are either incoherent or do not represent the context of the sentence. As an example, the excerpt “*excellent reception and crisp voice quality*” misrepresents the statement “*My previous phone had excellent reception and crisp voice quality*”. False positives of this kind are a result of errors in excerpt extraction. Excerpts that misrepresent the original statement can be a consequence of a comparison to another product, usage of grammatical clauses such as contrastive conjuncts such as *however* or *but*, or the usage of negation. Some of these are known to be challenging problems in natural language. Further, while we have used 8 words as an upper bound on the excerpt length for the purpose of illustration, this can be varied based on the underlying application. Variation in the upper bound will affect the quality of the blurbs generated and the effectiveness of the ranking orders produced. Although, that is out of the context of this work, analysis of the impact of an upper bound on the length of blurb can be further explored.

Product	Aspect	Review Sentence	Blurb
Video game	storyline	Blood Omen: Legacy of Kain is awesome with great graphics and a challenging <i>storyline</i> .	great graphics and a challenging storyline
GPS unit	screen	It is a great unit and is very easy to operate and the <i>screen</i> is very easy to see and work.	the screen is very easy to see and work
Drill	weight	This one is a powerful light <i>weight</i> cordless drill.	powerful light weight cordless drill
Book	advice	It has good simple truthful <i>advice</i> , with a down to earth feel, anyone can follow this book from start to finish.	good simple truthful advice

Table 1: Illustration of blurb generation for example review sentences. Here, the corresponding product-specific aspect and the extracted excerpt for the given review sentence are shown.

Figure 6: Plots showing the (a) Precision@k and (b) NDCG@k values computed for the evaluation dataset. Here k is the number of top blurbs considered for NDCG computation.

5 SUMMARY

The practice of using blurbs attributed to expert or critic reviews, to capture the interest of audience and promote a product has been pervasive in the field of books and movies for a long time. This concept is gaining increasing adoption in other domains such as consumer product websites where excerpts from expert reviews are used to encourage shoppers to explore a product. With the growing prominence of product reviews on most modern e-commerce platforms and the increasing reliance of online shoppers on reviews, it seems natural to introduce blurbs to e-commerce products. In the context of e-commerce product reviews, blurbs can simplify shopping experience when facing a myriad of product choices, in addition to highlighting interesting reviewer opinions on a product.

Traditionally, blurbs have been carefully handpicked by book editors or movie publishers. However, owing to the scale of modern e-commerce websites with hundreds of millions of product and product reviews, automated methods of discovering blurbs must be developed. While researchers in the area of text mining have proposed several methods focused on review mining, specifically assisting in the tasks of opinion mining, sentiment analysis and summarization, blurb mining from product reviews has not been explored. Although blurb mining has several commonalities with such methods, this task differs in several key areas. Blurb mining involves identifying product-specific excerpts, and the focus is on identifying short excerpts that retain the context rather than complete sentences. Further, the goal in blurb mining is to identify

interesting and enticing excerpts rather than summarizing the reviews. Moreover, it is critical to maintain user trust on e-commerce platforms to retain context of the opinions and not misrepresent them while identifying excerpts from sentences. These unique challenges, coupled with the scale and diversity of product catalogues on e-commerce websites necessitate developing focused methods for blurb mining.

In this paper, we introduced the problem of blurb mining from e-commerce product reviews, and proposed a three-phased approach to discover interesting blurbs. Specifically, we proposed methods to first discover product-specific aspects in reviews, next generate excerpts from qualifying review sentences and then rank them to select the interesting ones. We evaluated the quality of the extracted blurbs and the ranking order produced using precision@k and NDCG@k measures respectively. Although the focus of this work is on introducing the problem of blurb mining from e-commerce product reviews and presenting an approach to implement blurb discovery and ranking at scale, there are several avenues for further exploration and improvement. First, the problem of excerpts misrepresenting original sentences can be further explored, specifically with emphasis on comparative statements and sentences with contrastive clauses and negation. While there is a body of research that exists in the area of aspect mining, further research into integrating aspect mining into blurb discovery can be valuable. In this work, we employed several features pertaining to aspect-specific metrics in reviews and features that quantify sentiment to assist the classifier in distinguishing excerpts that are interesting against those that aren't. An excerpt being interesting, or in other words

the catchiness of a blurb is a very subjective concept and can be difficult to quantify. Further research in quantifying the catchiness of a blurb can be beneficial.

REFERENCES

- [1] Muhammad Abulaish, Mohammad Doja, and Tanvir Ahmad. 2009. Feature and opinion mining for customer review summarization. *Pattern recognition and machine intelligence* (2009), 219–224.
- [2] Henry Alford. 2007. Literary Misblurbings. (April 2007). Retrieved April 29, 2007 from <http://www.nytimes.com/2007/04/29/books/review/Alford.t.html>
- [3] Jinho D Choi, Joel R Tetreault, and Amanda Stent. 2015. It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool.. In *ACL (1)*. 387–396.
- [4] Marie-Catherine De Marneffe and Christopher D Manning. 2008. *Stanford typed dependencies manual*. Technical Report. Technical report, Stanford University.
- [5] Colin Dwyer. 2015. Forget The Book, Have You Read This Irresistible Story On Blurbs? (Sept. 2015). Retrieved September 27, 2015 from <https://www.npr.org/2015/09/27/429723002/forget-the-book-have-you-read-this-irresistible-story-on-blurbs>
- [6] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J Simske, and Luciano Favaro. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications* 40, 14 (2013), 5755–5764.
- [7] Mª Gea Valor. 2005. Advertising books: A linguistic analysis of blurbs. *Iberica* 10 (2005).
- [8] Maria-Lluïsa Gea-Valor and Marta Inigo Ros. 2009. On the dynamic nature of genre: A diachronic study of blurbs. In *Academic Evaluation*. Springer, 199–216.
- [9] Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1373–1378. <https://aclweb.org/anthology/D/D15/D15-1162>
- [10] Mingqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, Vol. 4. 755–760.
- [11] Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1035–1045.
- [12] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 415–463.
- [13] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated Rule Selection for Aspect Extraction in Opinion Mining.. In *IJCAI*. 1291–1297.
- [14] Wencan Luo, Fei Liu, Zitao Liu, and Diane Litman. 2016. Automatic summarization of student course feedback. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 80–85.
- [15] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093–1113.
- [16] Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of biomedical informatics* 52 (2014), 457–467.
- [17] Samaneh Moghaddam and Martin Ester. 2011. ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 665–674.
- [18] Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 339–348.
- [19] Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data* (2012), 43–76.
- [20] Patrick Nguyen, Milind Mahajan, and Geoffrey Zweig. 2007. Summarization of multiple user reviews in the restaurant domain. *Microsoft Research, Redmond, WA, CiteSeer* (2007).
- [21] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108 (2016), 42–49.
- [22] Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. 2014. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*. 28–37.
- [23] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics* 37, 1 (2011), 9–27.
- [24] Huifeng Tang, Songbo Tan, and Xueqi Cheng. 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications* 36, 7 (2009), 10760–10773.
- [25] Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 111–120.
- [26] Ivan Titov and Ryan T McDonald. 2008. A Joint Model of Text and Aspect Ratings for Sentiment Summarization.. In *ACL*, Vol. 8. 308–316.
- [27] Takaaki Tsunoda, Takashi Inui, and Satoshi Sekine. 2015. Utilizing review analysis to suggest product advertisement improvements. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 41–50.
- [28] Tao Wang, Yi Cai, Ho-fung Leung, Raymond YK Lau, Qing Li, and Huaqing Min. 2014. Product aspect extraction supervised with online domain knowledge. *Knowledge-Based Systems* 71 (2014), 86–100.
- [29] Zhijun Yan, Meiming Xing, Dongsong Zhang, and Baizhang Ma. 2015. EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management* 52, 7 (2015), 850–858.
- [30] Rong Zhang, Wenzhe Yu, Chaofeng Sha, Xiaofeng He, and Aoying Zhou. 2015. Product-oriented review summarization and scoring. *Frontiers of Computer Science* 9, 2 (2015), 210–223.