# A Study on Video Game Review Summarization

**George Panagiotopoulos**
Dpt. of Informatics & Telecommunications
University of Athens
Athens, Greece
`sdi1400136@di.uoa.gr`

**George Giannakopoulos**
NCSR Demokritos
Athens, Greece
`ggianna@iit.demokritos.gr`

**Antonios Liapis**
Institute of Digital Games
University of Malta
Msida, Malta
`antonios.liapis@um.edu.mt`

## Abstract

Game reviews have constituted a unique means of interaction between players and companies for many years. The dynamics appearing through online publishing have significantly grown the number of comments per game, giving rise to very interesting communities. The growth has, in turn, led to a difficulty in dealing with the volume and varying quality of the comments as a source of information. This work studies whether and how game reviews can be summarized, based on the notions pre-existing in aspect-based summarization and sentiment analysis. The work provides suggested pipeline of analysis, also offering preliminary findings on whether aspects detected in a set of comments can be consistently evaluated by human users.

## 1 Introduction

The rapid growth of video game industry with new products and technology has significantly increased the popularity of video games. As video games have now become one of the most profitable source of entertainment worldwide, the competition between development companies has increased notably.

Catering for gamers' needs is a demanding task that developers struggle to deal with. Thus, it is crucial for game companies to understand the overall consensus about their products. Additionally, what other people think of a game can also be an important piece of information for potential buyers. Video game reviews offer user-generated data that can be processed in order to identify both people's concerns and user-perceived quality of the game. A number of publishers (Steam[1], GoG[2], etc.) offer a wide range of games, spanning various genres. By visiting such a publisher's store, people are able to look through a game's description and its features, delve into the reviews of the game provided by other users and experts, but also contribute their own review. As some of the games can have millions of reviews, the large scale of information poses the need and challenge of automatic summarization.

The aims of the present paper are:

- to examine if and how aspect-based summarization and sentiment analysis can be applied on the domain of game reviews

- to propose a first approach on game review summarization,

- to offer an evaluation process on the performance of the game review summarization task.

The rest of the paper is structured as follows. In Section 2 we overview the research endeavours related to this work, uniquely positioning it in the current research spectrum and discussing the unique setting of game review summarization. In Section 3, we formulate the problem of game review summarization. In Section 4 we propose

---

[1] `https://store.steampowered.com/`
[2] `https://www.gog.com/games`

an approach to tackle the problem at hand, while in Section 5 we validate the performance of our method in a user study. We conclude the paper with a summary of the findings and future work, in Section 6.

## 2 Related Work

The importance of analyzing user reviews has drawn a great deal of interest among researchers. There has been a plethora of studies presenting different approaches on sentiment analysis as well as summarization of user reviews from various domains, such as product reviews and movie reviews. In the following paragraphs we overview such approaches, trying to sketch the research landscape and position this work with respect to other works.

Turney (Turney, 2002) suggests a PMI-based approach for classifying reviews from four different domains (e.g. automobiles, movies, e.t.c) as *recommended* or *not recommended*. His approach consists of three main steps: phrase extraction from a given review by applying POS tagging, orientation estimation for each phrase based on the PMI score between the phrase and the words *excellent* and *poor*, review labelling based on the average orientation of its phrases. In (Hu and Liu, 2004) Hu and Liu present an approach for generating a feature-based opinion summary from a large number of reviews. They propose promising techniques for each stage of their method, which aims at classifying sentences rather than each review as a whole. They present, among others, an iterative algorithm for identifying the underlying sentiment of a word using a small set of seed adjectives combined with WordNet's synset relations (Miller, 1998).

Similarly, Zhuang et al. (Zhuang et al., 2006) propose their approach for producing feature-based summaries on the domain of movie reviews. They make use of regular expressions and WordNet for feature mining and opinion word identification respectively. POS-tag patterns are used in order to identify feature-opinion pairs. Their experiments produced lower precision and recall scores than the results obtained in the domain of product reviews (Hu and Liu, 2004), mainly because of the peculiarity of movie reviews. Instead of just producing an opinion summary, (Jmal and Faiz, 2013) assess the opinion strength on a product and its features, while exploiting Twitter posts to highlight the most relevant features more effec-

tively. In a more recent work (Rist et al., 2018) identify aspect-based statements from product reviews through patterns extracted from dependency parse trees.

A number of studies have proposed supervised learning approaches by training sentiment classifiers. Pang and Lee (Pang et al., 2002) attempt to classify movie reviews using Naive Bayes, SVM and Max Entropy and multiple feature combinations. Their results indicate that ML techniques on sentiment classification can achieve high accuracy when feature presence instead of feature frequency is used. In (Wilson et al., 2005) the authors attempt to recognize phrase-level contextual polarity by using a two-step process. They firstly classify expressions as polar or neutral and subsequently classify the polar ones as positive, negative or neutral.

A novel flexible summarization framework, called Opinosis, is proposed by Ganesan et al. in (Ganesan et al., 2010). It is a graph-based approach that represents review text as a graph with unique properties and identifies various paths in it, each one acting as a candidate summary. The SMACk system (Dragoni et al., 2018) is an argumentation-based opinion mining framework which detects and extracts aspects coupled with polarities from documents by creating an argumentation graph.

Topic modeling has been widely used as a basis to perform extraction and grouping of aspects. Titov and McDonald (Titov and McDonald, 2008) introduce a Multi-grain LDA model which models global topics and local topics that capture ratable aspects and properties of reviewed items respectively. Their method is particularly suited to aspect extraction from reviews as it does not only identify important terms but also clusters them into coherent groups. In (Lu et al., 2009) aspects in eBay's sellers feedback comments are discovered using PLSA-based techniques. The authors try to group aspect terms that tend to co-occur in comments. Jo and Oh (Jo and Oh, 2011) proposed two generative models to discover aspects and sentiment in reviews. Sentence-level LDA (SLDA) constrains that all words in a single sentence be drawn from one aspect. Aspect and Sentiment Unification Model (ASUM) unifies aspects and sentiment and discovers pairs of aspect, sentiment, which we call senti-aspects.

Recent advances in computing hardware to-

gether with the increased availability of data have led to the ubiquitous use of neural networks as an effective tool for producing summaries and identifying sentiment in text.

In (dos Santos and Gatti, 2014) the authors develop a deep convolutional neural network that exploits from character- to sentence-level information to perform sentiment analysis of short texts. Conversely, in (Severyn and Moschitti, 2015) the authors construct a network with just a single convolutional layer and also presented a new model for initializing the weights of the network. A novel deep learning approach to aspect extraction is shown in (Poria et al., 2016) where a 7-layer CNN is combined with linguistic patterns. Using the dataset made available by Pontiki et al. (Pontiki et al., 2016), the authors in (Ruder et al., 2016) propose a hierarchical LSTM-based approach for that task of aspect-based sentiment analysis whilst a Cascaded-CNN architecture is presented in (Wu et al., 2016).

Despite the widespread appeal of video games, there has been little discussion on the domain of game reviews. Yauris and Khodra (Yauris and Khodra, 2017) propose an aspect-based summarization system for Steam reviews. They employ a modified double propagation (DP) algorithm for extracting aspect-sentiment word pairs. Following this, they use a seed list and word similarity to categorize aspect terms into groups, thus producing an aspect-based summary. In (Baowaly et al., 2019) the authors developed a robust model using Gradient Boosting Machine algorithm to predict the Steam review helpfulness.

Most works so far have relied on supervised learning methods by utilizing annotated datasets (Pontiki et al., 2016). As there is currently no existing dataset for aspect-based game review summarization, our work is designed with the aim to minimize the role of supervision. Furthermore, in our undertaking we take into account the following idiosyncrasies of the game setting:

- The folksonomy (dynamic) nature of the terms used in comments. Each genre and possibly game appear to be mapped to specific expectations by its users and, consequently, aspects that the users comment on. There appears that a fixed ontology or aspect set would not be sufficient to describe the aspects of all the game genres that get published over time. This is further accentu-

ated by the fact that hybrid games, combining genres, become a common sight.

- The possible vagueness of aspects, based also on the above comment. We thus examine whether aspects identified through an unsupervised process can be consistently labeled by humans.

- The fact that it is important to hold not a single response of sentiment, but understand the full spectrum of sentiments of players. This means that a single "positive", "negative" or "neutral" answer to how people have commented for an aspect is only a secondary finding. The distribution of comments over these three labels is more interesting and useful, and may be the primary aim of a game review summarization process.

Given the above analysis, we establish a problem definition in the following paragraph, trying to formally frame the game review summarization problem.

## 3 Problem Definition

As noted, video game reviews are likely to discuss several aspects of the game, such as graphics, gameplay, community e.t.c. Expert/professional reviewers tend to follow specific patterns of summarizing reviews, utilizing the above established aspects. They also provide an overall recommendation and possibly grade, while oftentimes they highlight "pros" and "cons" of the reviewed game. These pros and cons essentially designate the specific, non-formalized, aspects of a game (and possibly of other games of its genre). On the other hand, we should note that the expert reviewers only summarize their own review, which forms a single-document setting. In our case, we examine an approach more suited for a multi-document summarization setting, where several texts (reviews) are to be summarized in a single summary.

To take into account the above "gold standard" human approach, while tackling the multi-document differentiation, we formulate the problem as follows:

Given a set of game reviews $R = \{r_1, r_2, ...\}$ for a game $g$, the game review summarization task tries to perform the following steps:

**aspect identification** identify the set $A$ of aspects of the game, that the reviews $R$ comment on.

**aspect labeling** map each aspect $A$ to a label set $L_A = \{l_1, l_2, ...\}$, where each of $l_i$ is a (possibly weighted) term.

**sentiment extraction** extract a sentiment distribution $S_A$ of the form $S_A = \{s_{\text{positive}}, s_{\text{neutral}}, s_{\text{negative}}\}$, which describes the user sentiment over each aspect $A$.

**highlight extraction** extract the subset $P \in A$ of "pros", where $s_{\text{positive}} > s_{\text{negative}}$ and the subset $C \in A$ of "cons", where $s_{\text{positive}} < s_{\text{negative}}$.

**review summary** generate a single summary $\mathbb{S}$ containing all the above information.

Within this work we focus on the *aspect identification*, *aspect labeling* steps. We also touch the *sentiment extraction* and *highlight extraction*, providing baseline implementations. In the following paragraphs, we elaborate on the suggested methods that implement these steps.

## 4 Proposed Method

In this section we describe in detail the steps of our proposed method. Given a game, we first fetch a set of reviews, which are subsequently split into sentences. After having processed each sentence, we represent them using a bag-of-words (BOW) model. Sentences are then clustered followed by sentiment analysis on each cluster. The individual processing steps are described below :

### 4.1 Review Representation

Text representation plays a major role in the effectiveness and accuracy of clustering algorithms (Aggarwal and Zhai, 2012). In our approach we represent each review as a set of processed sentences, which are then converted into tf-idf vectors. More precisely, after having segmented each review into sentences, tokenization, stopword removal and lemmatization are applied on each one. For this purpose we employ spaCy v2.0[3], an opensource software library for advanced NLP. Following this, we convert the sentences into tf-idf vectors. Below we provide a review excerpt followed by the extracted sentences:

*Global offensive is not the key evolution point that we were hoping for and the response from the*

---

[3]https://spacy.io/

| | |
|---|---|
| **Cluster 1** | story, character, mode, main, mission |
| **Cluster 2** | money, spend, earn, waste, real |
| **Cluster 3** | time, fun, long, loading, screen |
| **Cluster 4** | reason, ban, permanently, innocent, account |
| **Cluster 5** | support, great, bad, community, good |

Table 1: Most frequent words in each cluster

*community often reflects this view. It is still however a glorious experience that sets a benchmark for all multiplayer shooters.*

- global offensive key evolution point hope response community reflect view

- glorious experience set benchmark multiplayer shooter

We also examined whether a word embedding would provide better results. However, the BOW representation method appeared to give more coherent results in the clustering step. It is very likely that the short length of sentences combined with the large vocabulary size has led to this finding. Thus, capturing the context of each sentence via a sentence2vec method can be challenging, probably requiring more specific training data.

### 4.2 Aspect Extraction

In this step we try to extract the aspects of a game that are mainly discussed by the reviewers. Video game aspects can be either explicitly or implicitly mentioned in a review text. For example, the sentence "*Easily my favorite game with realistic graphics*" clearly expresses an opinion about the aspect "*graphics*". On the contrary, the sentence "*The grenade explosions are so fake*" does not mention the word "*graphics*" but it obviously refers to the graphics of the game, or possibly the physics engine.

We apply k-means clustering on the previously collected sentences with the aim of producing a cluster-wise summary. The intuition behind this approach is that the produced clusters will exhibit the most salient aspects appearing in the reviews. In Section 5 we elaborate on our decision regarding the number of clusters.

Table 1 lists the most frequent terms appearing in each cluster. As anticipated, the words are semantically close to each other and they seem to represent a specific game aspect. We choose these terms to label the aspect cluster.
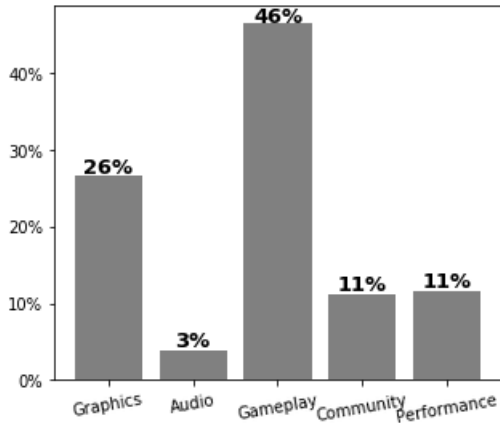
Figure 1: Aspect proportions exhibited in cluster on "The Elder Scrolls V: Skyrim" game

### 4.3 Aspect Labeling

Another way to label the aspects is to map them to a predefined set of aspect labels, based on gold-standard (i.e. professional) reviews. In Table 2 we show an indicative, human-provided mapping between terms and predefined aspect labels.

Based on the mapping illustrated in Table 2, each sentence can be classified into one of the aspects, by identifying the prevalent aspect of the sentence's words (i.e. terms). For instance, if the majority of the terms in a sentence belong to the *community* aspect, then the sentence is given this label. It should be noted that the term lists needs to be slightly modified based on the game's genre. The reason is that the terms that illustrate the "*gameplay*" aspect of a first-person shooter game differ notably from those of a puzzle or an adventure game. This fact highlights the intricacies of the game review task, where secondary (latent) variables alter the aspect descriptions.

Eventually, we end up with the predefined aspect proportions that each cluster exhibits. In Figure 3 we provide a few indicative sentences from a specific aspect cluster. Then, in Figure 1 we show how the sentences of the cluster led to a distribution over the predefined aspects.

### 4.4 Sentiment Analysis

The sentiment analysis step focuses on identifying the underlying sentiment that pervades each cluster. Since our clusters consist of sentences we perform sentence-level sentiment analysis.

As there is no sentiment analysis dataset specific to our domain, we decided to use VADER (Hutto and Gilbert, 2014), a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. Interestingly, VADER can tell us how positive, negative and neutral a given sentence is, instead of just classifying the sentence in a single category. VADER combines a dictionary of lexical features to sentiment scores with a set of five heuristics (e.g. punctuation, degree modifiers, e.t.c). Consequently, by calculating the three sentiment scores for each sentence in a cluster and averaging, we can get the distribution of the reviewers' sentiment for this cluster.

### 4.5 Final Output

The final output of the process is an aspect-based summary of a set of reviews of a specific game, coupled with positive, negative and neutral sentiment proportions for each aspect. While the initial set of numerous unstructured reviews are considerably difficult to deal with, this type of summary enables the reader to retrieve the most relevant information about the game according to his/her need.

## 5 Experiments

### 5.1 Experimental Setup

For our experiments we used the Steam review dataset gathered by Zuo (2018). It consists of more than 7 million reviews obtained via Steam's API. Each review text comes with a plethora of features concerning both the game being reviewed and the reviewer. For our experiments, we only utilized the game's ID, the review itself and the number of "helpful" votes the review has received by other community members.

In our experiments, to speedup the clustering process we use only a sample of the reviews of each game consisting of the 10,000 most voted reviews. Due to the syntactical peculiarities found in user-generated reviews, we also had to perform some extra pre-processing together with the sentence segmentation. This involved repeating phrases and multiple whitespace removal as well as filtering out terms consisting of non-ASCII characters.

Moving on to the clustering process, as reported previously, we decided to use the k-means method. We settled for this method because of the high dimensionality of our data, making a hierarchical

| Graphics | graphic, visual, look, aesthetic, animation, frame |
|---|---|
| **Gameplay** | mission, item, map, weapon, mode, multiplayer |
| **Audio** | audio, sound, music, soundtrack, melody |
| **Community** | community, support, toxic, friendly, |
| **Performance** | server, bug, connection, lag, latency, ping, crash, glitch |

Table 2: Selected terms for each aspect

```
You get attached to so
many characters and the
world is amazing.

Probably the best open-world
rpg out there.

The vast open world is
absolutely stunning.

If you're looking for a way to
waste massive amounts of time
just trolling around a world
play this.
```

Table 3: Indicative sentences from cluster on "The Elder Scrolls V: Skyrim" game

approach computationally expensive.

Identifying the appropriate number of clusters $k$ can sometimes be one of trickier tasks in the study of clustering. In order to deal with this issue, we attempted to use the elbow method and we also performed Silhouette analysis (Rousseeuw, 1987). Nonetheless, no appreciably optimal k was found by the two methods. However, this is not particularly surprising, in light of the fact that the reviews address a wide range of themes. Thus, the more clusters we create, the higher the coherence will be. Considering, though, that we aim to produce a digestible aspect-based summary using these clusters, it would be irrational to produce too many of them. For this reason, we decided to work with 5 clusters.

## 5.2 Results

In order to reach a sound conclusion we have performed an empirical evaluation with four different human evaluators. Before describing our evaluation process, we remind the reader that we seek to provide an evaluation process for game review summarization. Given this requirement, we assess the coherence of the generated clusters and examine whether they can be mapped to specific game aspects in a consistent way by humans. This study allows us to understand whether steps of the problem, as formulated in Section 3,

can be evaluated consistently. For the final output of the whole summarization pipeline we expect that standard summary evaluation methods, such as MeMoG (Giannakopoulos and Karkaletsis, 2013) and ROUGE (Lin, 2004) will be useful.

We asked the help of 4 evaluators, who were fluent in the English language. The evaluators were given a set of 20 sentences fetched from each of the five clusters of three different games (for a total of 15 clusters). We also opted for different genres in order to examine the inter-genre differences with respect to the terms used for describing game aspects. They were then asked read each set of sentences and complete the following tasks:

- Select up to $n$ representative sentences from the aspect cluster to represent/summarize the cluster. The idea behind this task is to show whether the cluster was coherent enough to be described by a representative subset of its sentences. The lower the number of representative sentences one would need to use to represent the cluster, the higher the coherence of the cluster.

- Describe the theme of each set using 3 to 5 (possibly multi-word) terms. This task aims to see whether humans can consistently label a given aspect cluster. If so, then the agreed wording(s) can be considered gold-standard, similarly to a Pyramid evaluation (Nenkova and Passonneau, 2004).

- Select one or more predefined terms (gameplay, graphics, audio, community, performance, overall, other) that best describe the aspect, according to the opinion of the human evaluator. We also allowed the user to select "other" as an option, to examine whether a significant number of aspects go beyond the predefined ones. This would indeed indicate the dynamic nature of aspects in the game review summarization setting.

In the "select representative sentences" task, we quantify how many sentences on average were selected by the evaluators to represent the cluster.

| ClusterID | Mean | +/- Std. Err |
|---|---|---|
| 0 | 9.75 | 2.14 |
| 1 | 7.00 | 1.68 |
| 2 | 9.00 | 2.42 |
| 3 | 7.50 | 2.02 |
| 4 | 5.25 | 1.80 |
| 5 | 9.25 | 2.17 |
| 6 | 5.00 | 1.73 |
| 7 | 5.75 | 1.25 |
| 8 | 7.00 | 1.22 |
| 9 | 4.25 | 1.60 |
| 10 | 8.50 | 2.78 |
| 11 | 4.25 | 1.31 |
| 12 | 5.00 | 1.08 |
| 13 | 9.00 | 1.47 |
| 14 | 8.00 | 2.48 |

Table 4: Average representative sentences per aspect

We expect that the lower the number, the better the coherence of the cluster. In Table 4 we see, for each cluster, the average number of sentences selected as representative by the users, plus the standard error. We see that, given 20 sentences, the users selected on average from 4 to 9 representative sentences.

In the "describe the theme" task, we examine whether humans can assign consistent labels in an open terminology setting (i.e. without limiting the possible labels). To measure the agreement here we post-processed their terms, semi-automatically creating equivalence classes of terms (which could also have been determined based on an embedding or a linguistic resource). Indicative equivalence classes were:

- ban; ban possible; bans

- best game; best rally game; buy; buy game; buying recommendations; described as best game; ...

- bad community; community; community bad; community sucks; low rank player behaviour bad; toxic community

We then examined, for each cluster, the number of equivalent terms that were used across all evaluators to label the specific aspect cluster. If at least 2 of the 4 evaluators utilize equivalent terms, we consider that the labeling is possible and successful. In all the 15 clusters at least one equivalence class was used consistently. In Figure 2 we show the consistently used equivalence classes per cluster. [4]

----

[4]There are cases where a single evaluator used more than

## 6 Conclusion

In this paper we discuss the domain of game review summarization. We highlighted main challenges of the domain, showing that a number of unique traits require different approaches from other summarization settings. We formally expressed a view of the task, suggesting a baseline implementation. We then described a possible evaluation process, aiming to quantify the success of the aspect identification and labeling, taking into account coherence and consistent labeling from human evaluators.

This preliminary study of the game review setting opens a number of research questions that we can pursue in the future. First, how does the game genre affect the aspects of a game? Is there a causal relation that connects them? Can we perform automatic evaluation with or without human gold standard summaries? What is different from other summarization settings, concerning the evaluation?

In this work, we offer a first research step towards the emerging and useful domain of game review summarization. We understand that this first step simply highlights interesting points of focus, while providing some intuition on what is meaningful and doable from an evaluation perspective. We feel confident that this will help document and formulate a consistent setting and benchmarking process, helping related endeavors grow in the future.

## References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining Text Data*.

Mrinal Kanti Baowaly, Yi-Pei Tu, and Kuan-Ta Chen. 2019. Predicting the helpfulness of game reviews: A case study on the Steam store. *Journal of Intelligent & Fuzzy Systems* 36(5):4731–4742.

Cicero dos Santos and Maira Gatti. 2014. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 69–78.

----

one term from an equivalence class, thus leading to counts over 4. However, our counting algorithm ascertains that at least 2 different evaluators will have used a term from the same equivalence class, before increasing the count.
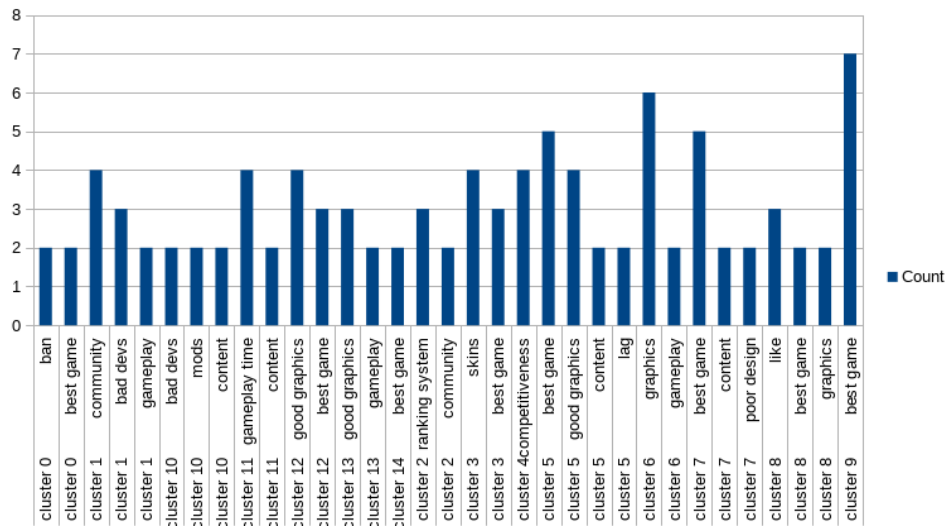
Figure 2: Count of usage for consistent equivalence classes per cluster

Mauro Dragoni, Celia da Costa Pereira, Andrea G. B. Tettamanzi, and Serena Villata. 2018. Combining argumentation and aspect-based opinion mining: The SMACk system. *AI Communications* 31(1):75–95.

Kavita Ganesan, Chengxiang Zhai, and Jiawei Han. 2010. Opinosis: a graphbased approach to abstractive summarization of highly redudant opinions. In *In COLING*.

George Giannakopoulos and Vangelis Karkaletsis. 2013. Summary evaluation: Together we stand npower-ed. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 436–450.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*. volume 4, pages 755–760.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Jihene Jmal and Rim Faiz. 2013. Customer review summarization approach using Twitter and SentiWordNet. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13*. ACM Press, Madrid, Spain, page 1.

Yohan Jo and Alice H. Oh. 2011. Aspect and Sentiment Unification Model for Online Review Analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, WSDM '11, pages 815–824. Event-place: Hong Kong, China.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. pages 74–81.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated Aspect Summarization of Short Comments. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '09, pages 131–140. Event-place: Madrid, Spain.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*. pages 145–152.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphe De Clercq, Vronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud Mara Jimnez-Zafra, and Glen Eryiit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 19–30.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108:42–49.

Michael Rist, Ahmet Aker, and Norbert Fuhr. 2018. Towards Making Sense of Online Reviews Based on

Statement Extraction. In *7th International Conference on Natural Language Processing*. pages 01–12.

Peter Rousseeuw. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* 20(1):53–65.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. *arXiv:1609.02745 [cs]* ArXiv: 1609.02745.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '15, pages 959–962. Event-place: Santiago, Chile.

Ivan Titov and Ryan McDonald. 2008. Modeling Online Reviews with Multi-grain Topic Models. *arXiv:0801.1063 [cs]* ArXiv: 0801.1063.

Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *arXiv:cs/0212032* ArXiv: cs/0212032.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Haibing Wu, Yiwei Gu, Shangdi Sun, and Xiaodong Gu. 2016. Aspect-based Opinion Summarization with Convolutional Neural Networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Vancouver, BC, Canada, pages 3157–3163.

K. Yauris and M. L. Khodra. 2017. Aspect-based summarization for game review using double propagation. In *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*. pages 1–6.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie Review Mining and Summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '06, pages 43–50.

Zhen Zuo. 2018. Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier. Technical report, University of Illinois at UrbanaChampaign.