Summarizing Game Reviews: First Contact

Aris Kosmopoulos
SciFY PNPC and NCSR Demokritos
Athens, Greece
akosmo@scify.org

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

George Giannakopoulos
SciFY PNPC and NCSR Demokritos
Athens, Greece
ggianna@iit.demokritos.gr

Nikiforos Pittaras
NCSR Demokritos and Kapodistrian University of Athens
Athens, Greece
pittarasnikif@iit.demokritos.gr

ABSTRACT

In recent years the number of players that are willing to submit a video game review has increased drastically. This is due to a combination of factors such as the raw increase of video gamers and the wide use of gaming platforms that facilitate the review submission process. The vast data produced by reviewers make extracting actionable knowledge difficult, both for companies and other players, especially if the extraction is to be completed in a timely and efficient manner. In this paper we experiment with a game review summarization pipeline that aims to automatically produce review summaries through aspect identification and sentiment analysis. We build upon early experiments on the feasibility of evaluation for the task, designing and performing the first evaluation of its kind. Thus, we apply variants of a main analysis pipeline on an appropriate dataset, studying the results to better understand possible future directions. To this end, we propose and implement an evaluation procedure regarding the produced summaries, creating a benchmark setting for future works on game review summarization.

CCS CONCEPTS

• **Applied computing** → **Computer games**; • **Computing methodologies** → **Information extraction**.

KEYWORDS

summarization, natural language processing, sentiment analysis, game reviews, Steam.

ACM Reference Format:
Aris Kosmopoulos, Antonios Liapis, George Giannakopoulos, and Nikiforos Pittaras. 2018. Summarizing Game Reviews: First Contact. In *SETN2020: Games and AI, September 2–4, 2020, Athens*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The ever-expanding popularity of digital games is evidenced by the large profit margins of the commercial game industry sector [4], the vast and diverse swathes of the population that play games [17], and the appeal of games and gamification beyond the purposes of entertainment [13]. A large factor for the market penetration of digital games are distribution platforms such as *Steam* and the *Google Play Store*. Not only do these distribution platforms allow interested players to purchase and download new games, they also cultivate a player community with players returning to rate and comment on their favorite game or even contribute user-created content, strategies, cheats, etc. This community-driven content often informs other users' purchases (e.g. via an aggregated review score) but is also carefully monitored by developers and publishers in order to gauge opinions on specific aspects of the game which can be patched or improved in updates to the game or in sequels. For both players and developers, being able to succinctly monitor other players' views is highly beneficial. The website www.metacritic.com aggregates reviews by players and professional critics, returning a percentage score for the game and highlighting diverse reviews along the spectrum of positive versus negative. The Steam platform also aggregates its users' reviews into different categories ('Mixed', 'Overwhelmingly Positive', 'Mostly Negative' etc.) which is another criterion for sorting and (likely) promoting games. The simple aggregation of reviews into a general score is important, but it obfuscates the nuances of the different reviewers' grievances and is of limited use to designers who wish to improve their game. This paper explores techniques for text summarization in order to provide a multi-dimensional and holistic summary of Steam reviews for a particular game.

We explore the topic of summarization for game reviews using a large dataset of Steam reviews from 12 selected games. The goal of the summarization pipeline is to extract users' views on different facets of games such as graphics, audio, and gameplay [28], leveraging textual sentiment analysis to identify and positive and negative review snippets, creating a composite summary of indicative comments on a specific game facet. Unlike the numerical aggregation of Metacritic or Steam, this approach extracts individual sentences (and criticisms) contained within a usually dense review and attempts to classify those in terms of positive or negative automatically (rather than based on the user's binary recommendation). The presentation of the game's summary, which is split based on different aspects typically criticized in games, can be valuable for both players and designers. For players, the statistics derived from this process (e.g. ratio of positive versus negative comments in one aspect) can act as an expanded game scoring system not unlike professional game reviews which gave a score to graphics, audio etc. For designers, the indicative comments split per sentiment and aspect allows for a quick monitoring of players' current favorite features. Moreover, the flexible way in which aspects

SETN2020, September 2–4, 2020, Athens

ACM ISBN 978-1-4503-8878-8…$15.00
https://doi.org/10.1145/1122445.1122456

are defined allows designers to explicitly redefine the keywords they are interested in, personalizing the summary to their design priorities.

There has been very limited attention to game review summarization, besides student projects [55]. Inspired by the only work that performs aspect-based game review summarization [50], this paper evaluates the outcomes of a straightforward summarization pipeline in a small-scale user survey. Using the twelve most reviewed games in a 2017 dataset of Steam reviews, the resulting summaries are evaluated by a small set of experts. The paper studies pipeline variants to better sketch what is important in game review summarization. Based on the outcomes of the different summarization processes, and a small-scale study where the different outcomes were compared, a number of potential improvements were identified. The paper also highlights the many directions which game review summarization research can follow so that it can serve designers and players through different pipeline implementations, alternative visualizations, bottom-up aspect discovery, or text processing driven by domain knowledge.

The paper is structured as follows. We start with a review of related works in Section 2. We then describe the proposed summarization pipeline and variants in Section 3. We describe the dataset in Section 4 and present two different user studies in Sections 5 and 6. We then discuss the results in Section 7 and conclude the paper in Section 8.

2 RELATED WORK

User reviews are a rich source of information, although the extraction and analysis of this information can be challenging not only due to the textual nature of the medium but also because users tend to have a mixed opinion about various features [31]. Approaches such as sentiment analysis as well as summarization have been applied to various datasets, such as product reviews [24, 31], movie reviews [53, 54], or hotel reviews [25]. Section 2.1 surveys relevant approaches for the different phases of a summarization pipeline, while Section 2.2 discusses the nuances of the Steam platform and early work in game review summarization. For interested readers, [25] provides a more thorough overview on review summarization according to the type of corpora used as input.

2.1 Summarization Pipeline

Summarization can be *extractive* when relevant portions (usually sentences) of the input are copied and combined, or *abstractive* when new text is generated to rephrase and summarize the input [18]. The summarization pipeline requires a number of steps before the raw textual input can produce a summary; algorithms and approaches for each step are discussed below.

2.1.1 Pre-processing and parsing. A fundamental step towards summarization (and natural language processing more broadly) is the pre-processing and extraction of features from the dataset. In the analysis below, the term documents is used to describe any type of text, e.g. a sentence, a paragraph, or an academic paper. One popular if naive approach for pre-processing data is the *bag-of-words* which collects all words in the document, disregarding their order and grammar. This method counts the number of instances of the same word, and the frequency of occurrence of each word is used as a feature to measure similarity between documents. Since many words (such as articles or pronouns) are far more frequent in all documents, terms are weighted based on their frequency via *tf.idf* [41] where the term frequency (tf) is multiplied by the inverse document frequence (idf). Unlike the bag-of-words approach, the word order is considered in many other approaches as it can capture a word's importance. For instance, the first and last sentences in a larger document tend to be more important [33]. Other approaches tag words on their part-of-speech (POS) [37], e.g. nouns (NN), verbs (VB), or adverbs (RB). This is useful for pre-processing, e.g. selecting only sentences with a noun and adjective as a corpus for review summarization [25]. Another use of POS tags is to select N-grams (i.e. a sequence of words) with specific parts of speech, such as a comparative adverb followed by an adjective [47].

2.1.2 Topic Modeling. Identifying the topic of a document, sentence, or review is often necessary for clustering opinions on the topic together. When the topics of interest are known in advance, experts usually provide the keywords used to filter the relevant documents. For instance, TweetElect used an initial set of 38 keywords related to the 2016 US elections (including candidates' names) for streaming relevant tweets [11]. However, a boolean check whether a keyword is specifically mentioned is rarely sufficient due to the nuances of language; *query expansion* is applied to create a larger set of terms related to each original keyword [29]. Supervised learning is often applied for topic modelling, showing positive and negative examples of relevant documents to a classifier [29]. When topics are unknown and must be discovered from the data, a simple approach is to identify the most frequent terms and cluster emergent terms based on co-occurrence [16]. Probabilistic topic models such as Latent Dirichlet Allocation (LDA) [8] can more efficiently discover topics without domain knowledge, following a bag-of-words approach which disregards word or document order. LDA randomly chooses a set of topics and decomposes the probability distribution matrix of words in a document into two matrices consisting of the distribution of topics in a document and the distribution of words in a topic. Due to the vast number of possible topic structures, sampling-based algorithms are used to find the sample topics which best approximate the posterior distribution [7]. LDA has often been applied to find topics within reviews, primarily in order to identify review's sentiments towards these topics, e.g. in [26].

2.1.3 Sentiment Analysis. The sentiment behind utterances is important for summarization, especially when the corpus is reviews of any kind. Turney [47] highlighted that reviews may recommend or not a certain product, movie, or travel destination; a summary therefore should account for both positive and negative reviews. Turney's study was the first to perform sentiment analysis on text-based reviews based on responses of the AltaVista internet search query on how near the phrases were to the word 'excellent' (for recommended) and the word 'poor' (for not recommended). Manually created lexica for words that express sentiment have been used in conjunction with fuzzy logic, vector distance, etc. to classify positive and negative [12, 45]. In the same context, there has been extensive work on extracting opinion words which express subjective opinions within sentences [49]. It has been found that subjective sentences are statistically correlated with the presence

of adjectives [49], and much research in product review summarization uses adjectives to determine sentiment polarity. For instance, Hu *et al.* [24] used a frequency-based algorithm to find relevant domain features, and then extracted nearby adjectives to such domain features. Using a labeled set of adjectives and expanding the initial set via WordNet, Hu *et al.* classified the extracted adjectives' polarity and assigned that positive or negative sentiment to the nearby domain feature. The SentiWordNet database is constructed based on the same principles of the domain-specific adjective classification of [24], using a manually annotated set of seed words and using WordNet term relationships to expand the training set, which is then used as the ground truth for machine learning classifiers [2]. SentiWordNet, and similar general-purpose models for sentiment prediction [46], have been used for polarity detection in reviews, e.g. in [23, 40].

2.2 Steam Review Summarization

Since its 2003 release, the Steam platform has become the largest digital distribution platform for PC gaming [15], hosting over 34,000 games and tens of millions of active users daily. This paper focuses on user-created reviews on Steam, although other initiatives such as the Steam workshop allow users to upload their mods or strategies and comment on others' content. User reviews can be submitted only by people that have purchased the game from Steam, although they are visible to all. As noted above, Steam aggregates user reviews into a category and provides a number of companion statistics, including a timeline of reviewer's scores. Reviews themselves consist of a single binary recommendation (Recommended versus Not Recommended) and a text explaining the user's opinion. Other users can review the quality of the review itself by tagging it helpful, not helpful, funny, or breaking the Rules of Conduct. By default, Steam shows the most helpful reviews submitted within the last 30 days, although users can also choose to sort reviews by other criteria.

As noted in the introduction, there is no systematic academic research in Steam review summarization. To the best of our knowledge, the only academic publication that tackles the problem of aspect-based summarization on such data is by Yauris and Khodra [50]. In their approach, only relevant portions of sentences were extracted via conditions applied on text tagged via Parts of Speech; these portions were usually small, e.g. the phrase could be "amount of content" [50]. Similar to our approach, a pre-specified set of keywords are used for aspect categorization. The aspects and keywords are similar but not identical to our approach (e.g. the aspects in [50] are gameplay, story, graphic, music, community, and general/others), while choosing the aspect described in the phrase was based on the cosine similarity from each word of the phrase to the aspect's keywords. The output summary consists of many aspects (most of which are outside the pre-specified keywords) and a single adjective for each, unlike our current work which extracts complete sentences with different polarities. The summarization pipeline was tested on a single game (Skyrim), exploring different sentiment extraction approaches using precision and recall as performance metrics. While our current work does not explore as many parameters for sentiment analysis, it is the first instance where game review summaries are evaluated by humans in a small-scale but thorough user study.

3 SUMMARIZATION PIPELINES

Figure 1 visualizes the main components of our pipeline:

Preprocessing which aims to prepare the input reviews for further analysis. This may imply cleaning, chunking text in snippets or sentences, Part-of-Speech tagging, and other similar tasks.

Aspect Identification which identifies interesting aspects (or topics) in the reviews. These topics may be expressed as a set of words, e.g. "visual, aesthetic, scenery" or "soundscape, audio experience, "sound effects".

Aspect Labeling which assigns clear, descriptive labels to the discovered aspects. E.g. "graphics", "audio".

Sentiment Analysis which gathers information related to the sentiment expressed within the reviews. This information may later be used to update the final summary appropriately. For example, one may need only positive views in the summary, or—most probably—a sampling of all the views, be they positive or negative.

Summary Creation which implies the process which, given all the information gathered in previous steps, forms and renders the final summary for the user.

Given the above pipeline, we implemented three different variants. The first two are based on keyword detection and *Clustering* (CL). The first variant does not do Sentiment Analysis, while the second one uses the full pipeline. The last one is another full pipe method based on *Deep Learning* (DL) that focuses on improving on Aspect Labeling and Summary Creation steps.

3.1 CL pipeline

During the preprocessing step, each review is split into sentences, each sentence is cleaned in order to create the basic elements on which the final summaries will be based. The cleaning process included of some character replacements so that each sentence could be presentable (e.x. starting with a capital letter and ending with a period) even if it originated from a larger sentence that was split during sentence splitting. Moreover, preprocessing prepared the lemmatized versions of the sentences which are used for aspect detection. In these lemmatized sentences, general stop words are removed. For all preprocessing steps, we used the default functions (and stop word lists) of the *nltk* Python library [5].

The aspect detection process is split into two parts: aspect identification and aspect labeling. Aspect identification splits sentences into sets that focus on a specific aspect while aspect labeling identifies this aspect in order to present it to the final review summary.

Our approach uses a predefined set of aspects, presented in Table 1. We selected these six aspects since they are well-established facets of games [28] and are popular dimensions within professional reviews.

A simple approach for aspect labeling is to use a dictionary of keywords per aspect as the ones presented in Table 1. In order to be able to include sentences even when they do not include the exact keywords, a k-means clustering is applied to all sentences to find clusters with similar text. Terms are weighted based on their

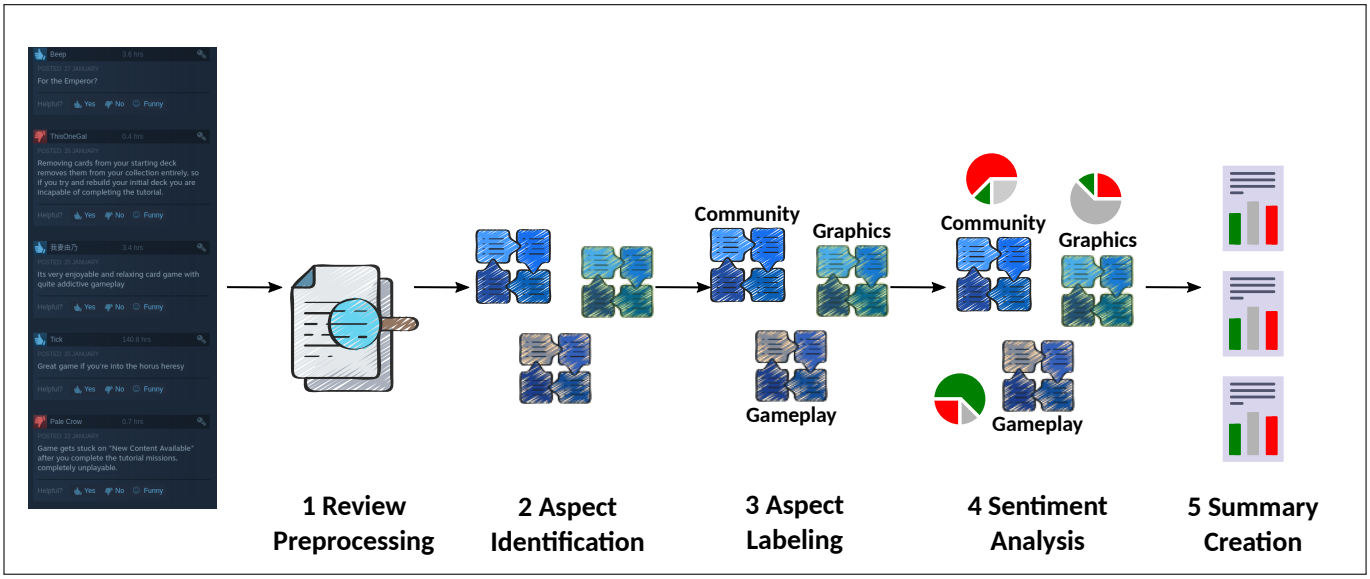

Figure 1: The full pipeline represents both the Clustering variant (CL Full) and the Deep Learning variant (DL Full), while variant CL AsDe produces summaries by skipping the Sentiment Analysis step.

Aspect	Keywords
Graphics	graphic, visual, aesthetic, animation, scenery
Gameplay	mission, item, map, weapon, mode, multiplayer, control
Audio	audio, sound, music, soundtrack, melody, voice
Community	community, toxic, friendly
Performance	server, bug, connection, lag, latency, ping, crash, glitch, optimization
Story	dialog, romance, ending, cutscene, story

Table 1: Aspects and keywords used for the identification of dominant aspects in review clusters.

frequency via $tf.idf$, which has been used extensively for sentence similarity in bag-of-words approaches (see Section 2.1). The result is K clusters of sentences with similar words to each other; in all our experiments we set $K = 20$ based on prior evidence [35]. Once sentences are all assigned a cluster based on the distance to the center, all sentences in all clusters are processed in the following fashion:

(1) If the sentence contains the exact keywords of only one aspect, the sentence is assigned to that aspect and is flagged as a candidate that can be used by the summary of that aspect.
(2) If keywords from multiple aspects are found in the sentence, the sentence is flagged as an unsuitable candidate for any summary and removed.
(3) If no aspect keywords are found in the sentence, the most common aspect within the sentences of the same cluster will be used to label this sentence and flag it as a candidate. For instance, if a sentence does not contain any keyword, but sentences in its cluster predominantly belong to the aspect *Gameplay* via case (1), then the sentence is also assigned to the same aspect and flagged as a candidate.

Using the sentences from cases (1) and (3), a set of candidate sentences is created per aspect. Using these sets, the first variation of our pipeline could now produce a summary. This variation, named Clustering Aspect Detection summary (CL AsDe), chooses N sentences at random from each aspect's set. A sample CL AsDe summary can be found in Table 2 for Tom Clancy's The Division.

The next step of the process is Sentiment Analysis, which is used by the next summarization variant (CL Full). Using the different sets of candidate sentences per aspect, the sentiment polarity (positive or negative) of each sentence is calculated by averaging the sentiment score of each word it contains. As above, sentiment analysis of each word is done via the default functions of the *nltk* Python library [5]. The library calculates probabilities for each polarity class (positive, neutral, negative). We took into account sentences which were assigned a class with a probability of at least 0.5. In order to select a number of sentences per category, a k-means clustering approach (using $tf.idf$) is applied within the set of sentences with the same polarity. In the CL Full implementation of this paper, only two sentences per polarity are selected ($k = 2$) as the ones closest to each cluster's centroid. If there exist sufficient positive and negative sentences, then this approach returns 6 sentences as bullet points. Note that if fewer than two sentences are above the threshold for positive (or below the threshold, for negative) then fewer sentences may be included in the summary. An example summary from CL Full variant can be found in Table 2 for Tom Clancy's The Division.

3.2 DL pipeline

After experimenting with the first two variant pipelines and taking into account the feedback of the first user study (see Section 5), we decided to focus on improving the following:

- Keyword detection and clustering based Aspect Labeling must be improved to avoid sentences such as "If those things all sound good to you you will like the game." to be labeled as audio sentences.
- The final summary should somehow provide information regarding the whole sentiment of the given aspect and not just by the selected sentences.
- The final summary should use a better sentence extraction approach in order to deal with redundancy.

Taking all the above into account, the DL pipeline makes changes to the Aspect Detection and Summary Creation steps of the CL pipeline described in Section 3.1.

For Aspect Detection, we used the BERT model [14] to generate embeddings for game reviews. BERT is a deep neural language model that uses a bidirectional, multilayer transformer architecture, exploiting cross and self-attention to capture word interdependencies effectively [3, 48]. The approach relies on multi-head attention modules for sequence encoding modelling, with word order information being retained with additive positional encoding vectors. BERT is trained in an unsupervised setting on large quantities of English text, using masked language modelling and next sentence prediction objectives. These tasks require the prediction of hidden sequence tokens and the generation of an entire sequence, given an input sequence (e.g. for tasks such as question-answering and text entailment, etc.). This pretraining scheme and architecture have been shown to perform exceptionally well for a variety of natural language understanding tasks.

To obtain the representation for a game review, we feed the text to the model using a sequence length of 16 tokens. We use the $BERT_{BASE}$ model variant, that produces 768-dimensional sequence embeddings, learned during training for classification purposes. The implementation and pretrained model utilized are provided by the transformers software package from *huggingface*[1]. Using the produced embeddings as features we trained a binary Ridge Logistic Regression classifier [19] (one vs all) for each aspect. We also trained a seventh classifier to detect sentences unfit for any aspect. For each candidate sentence a confidence score was calculated by each aspect classifier. Only sentences with a high prediction confidence in the given aspect and a low confidence on each other classifier were selected as summary candidates for the next steps of the pipeline.

During the Summary Creation we applied the following strategy to the 100 most probable candidate sentences of each aspect. First, the NewSum Toolkit [20] was used to select the sentences that provide the most representative information. NewSum uses language-agnostic methods based on n-gram graphs, that not only extract the most representative sentences, but also deal with redundancy. In the end we had 20 candidate sentences per Aspect. The final summary was composed by 6 sentences using the following strategy:

- Select the most positive sentence (Sentiment Analysis).
- Select the most negative sentence (Sentiment Analysis).
- Select the first 3 sentences provided by NewSum Tookit (excluding the previously selected sentences).
- Create an artificial sentence using the polarities provided by Sentiment Analysis of all the aspect sentences. The polarity of each sentence was mapped as 1, 0 or -1 (positive, neutral, negative) using thresholds. Given an Aspect and the mean Polarity score $\bar{P}$, the possible produced sentences reflect opinions that fall in the following categories:
  - **Mixed:** $\bar{P} \approx 0$, high standard deviation.
  - **Mostly neutral:** $\bar{P} \approx 0$, low standard deviation.
  - **Mostly positive:** $\bar{P} > 0$ above a threshold.
  - **Mostly negative:** $\bar{P} < 0$ below a threshold.

The final summary is composed by randomly shuffling these 6 sentences. An example summary from DL Full variant can be found in Table 2 for Tom Clancy's The Division.

[1] https://huggingface.co/

- In a few words the game is single dimensional this might sound vague but it becomes apparent that there is not much depth as you play once you're a couple hours in. - Clothes sound "right" when you move in them. - They sound good and looked good with ability to mod for better stats or even rerolling stats. - They have improved the pve portion of the game and crazy as it sounds the pvp too. - No music and something feels so strangely abandnded about it. - Like how if there's a blizzard your cap and shoulder will be covered in snow and that npc voices will echo when they are standing in hallways with hollow walls.
- Very good voice acting. - Great abilities pretty good sounds; indoor echos reverb off objects etc. - If those things all sound good to you you will like the game. - Superb voice acting and ambient city sounds are also a good plus for this game. - It sounds hyperbolic but I'm being dead serious. - Sounds terrible right
- Most opinions are positive regarding audio. - The voice acting in the game is in the higher tiers as is most ubisoft games. - There are not a lot of different voices and some of the voice acting for them is bad. - Ubisoft - bugs - the textures are so fucked up that nobody can play this game anymore. - And it clearly shows I want to play it and that I try to. -I'm gonna be honest the cinematics are pretty great.

Table 2: Summaries generated by different pipelines, for aspect *Audio* of Tom Clancy's The Division. From top to bottom: CL AsDe (only aspect detection), CL Full (aspect detection with sentiment analysis) and DL Full (Deep learning combined with a sophisticated summarizer).

4 DATASET

As a first demonstration of the summarization pipeline, we follow [35] and select the most helpful reviews on Steam, splitting them per game. This paper parses the Steam review dataset gathered by Zuo [55], which consists of over 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, although only a subset of features is used for this experiment. Since Steam users can vote a review as helpful, unhelpful, or spam, we only consider 'valid' reviews those with 10 or more user votes as 'helpful'. With this criterion (minimum of at least 1000 of 'helpful' reviews), we select twelve games with the most valid reviews (see Table 3). The games selected have a desirable diversity both in terms of genres (shooting, survival, adventure, open-world, multi-player, single-player, etc.) and in terms of general audience reception (shown by the Metacritic score which aggregates professional and users' reviews).

For each of the selected games we selected to keep the 10 thousand most up-voted reviews. As already discussed in Section 3 each

Game Title	Publisher	Year	Reviews	MC
No Man's Sky	Hello Games	2016	4146	61%
DayZ	Bohemia Interactive	2018	3349	–
PAYDAY 2	Starbreeze	2017	2573	79%
ARK: Survival Evolved	Studio Wildcard	2017	2368	70%
Grand Theft Auto V	Rockstar Games	2015	2104	96%
Firewatch	Campo Santo	2016	1599	81%
Darkest Dungeon	Red Hook Studios	2016	1564	84%
Just Survive	Daybreak Game Company	2015	1463	–
Killing Floor 2	Tripwire Interactive	2016	1276	75%
Elite Dangerous	Frontier Developments	2015	1270	80%
Tom Clancy's 'The Division'	Ubisoft	2016	1091	79%
Subnautica	Unknown Worlds Entertainment	2018	1056	87%

Table 3: Games selected from the dataset, sorted by the number of 'valid' reviews (10 or more 'helpful' votes). The Metacritic score (MC) is included for reference.

of these reviews was split into sentences to create a sentence pool per game. On average, the sentence pool consisted of around 50 thousand sentences per game. The smallest pool of sentences was for PAYDAY 2 (37K), while the largest one was for Elite Dangerous (70K). The average length of the sentences was 85.7 in characters and 16.4 in words. In terms of both characters and words, the longest sentences were those of Darkest Dungeon (average of 91.8 characters and 17.3 words) and the shortest ones were those of Just Survive (average of 79.9 characters and 15.6 words).

In terms of aspects, the most common one was *Gameplay* on average. *Performance* was the next most popular aspect and in certain games such as ARK: Survival Evolved it was the most popular one. The least popular aspect was *Audio* with a ratio of 1 to 5 compared to the *Gameplay* aspect.

In terms of sentiment, the majority of sentences were more neutral than positive or negative. Between positive and negative sentiment, no general safe conclusions can be drawn since the results varied given different combinations of aspects and games. In general, we can say that the aspect *Performance* was characterized as negative more frequently. The opposite was true for the aspect *Graphics*. On the other hand the sentiment ratio (positive vs negative) towards the aspect *Community* varied between different games.

5 FIRST USER STUDY

As a first experiment, we evaluated the two variations of the CL pipeline (CL AsDe and CL Full) in a small-scale user-study with summaries of aspects of the 12 games of Table 3.

5.1 Annotation Protocol

A pairwise comparison process was followed, rather than a scale-based rating approach, due to (a) evidence that comparison-based evaluation can be less demanding cognitively [9] and (b) a rich body

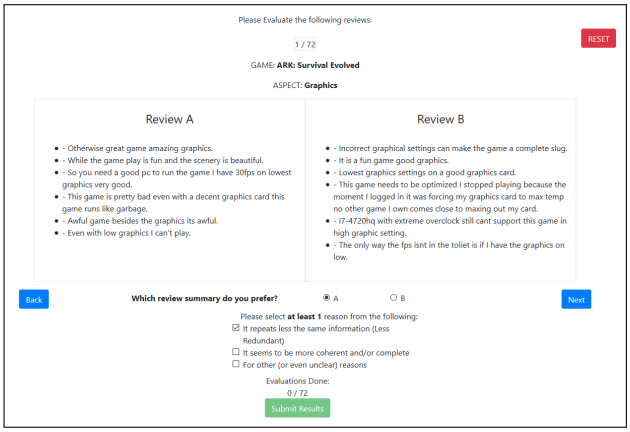

Figure 2: User interface for online evaluation of summaries produced by CL AsDe and CL Full methods.

of literature that has applied pairwise evaluation for summarization tasks [34] (e.g. the single document summarization task in [21]).

To this end, we created an online evaluation user interface (UI) (see Figure 2) which supported comparative pairwise evaluation of summaries. We initialized the system by providing two sets of summaries $\mathbb{A}, \mathbb{B}$, one from system A and one from system B. Each summary in $\mathbb{A}$ corresponded to a summary in $\mathbb{B}$, as they both summarize the same set of reviews and the same aspect (e.g. the aspect *Graphics* of DayZ). During the experiment, each system's summary was randomly placed first or second to minimize any bias related ordering effect.

The UI also informed the user of the title of the game being summarized, plus the aspect (e.g. *Graphics*). The user was then called to select their preferred summary (A or B) and explain the reasons for this preference. For the latter annotation, the user could select one or more tickboxes among the following options:

- It repeats less the same information (Less Redundant)
- It seems to be more coherent and/or complete
- For other (or even unclear) reasons

The first two options aim to assess whether redundancy is a concern and, similarly, whether coherence and completeness are useful in the task. Redundancy has been traditionally a summarization evaluation indicator [1], especially in multi-document summarization. The completeness and coherence aspect is essentially a (more nuanced) version of overall responsiveness, as this has been used in DUC/TAC summarization tracks and related work [10].

5.2 Participants

The evaluation was carried out by eight adult evaluators (3 female), fluent in English, with gaming experience. The evaluators were selected explicitly among the authors' network of contacts and invited directly by the authors. Participants were asked to connect to the online system and evaluate all 72 pairs of summaries (produced by CL AsDe and CL Full), which covered all predefined aspects (see Table 1) of the ten games in Table 3. There was no time limit for completing the evaluation, but there was a requirement that all pairs were evaluated in a single session.

Aspect	CL Full	CL AsDe
Audio	43%	57%
Community	51%	49%
Gameplay	55%	45%
Graphics	30%	70%
Performance	54%	46%
Story	49%	51%
Overall	47%	53%

Table 4: First user study: annotators' preference of one summarization algorithm over the other, per aspect and overall.

	Df	F value	p value
game	11	1.519	0.120
aspect	5	3.912	0.001 *
evaluator	6	7.945	0.000 *
coherence	1	18.6491	0.000 *
redundancy	1	5.7604	0.017 *
other	1	0.5639	0.453

Table 5: Analysis of variance between the preference of one approach and different factors. Significant findings are shown with an asterisk. The analysis is made on the F statistic and the degrees of freedom (Df) are also noted.

5.3 Results

The data collected from the experiment was a total of 576 observations, including the preference of each evaluator for each pair of summaries and the reasons for this choice. The primary goals of the user study are to assess (a) whether the annotators prefer one of the two summarization approaches, (b) which criteria they explicitly (via the three tickboxes) or implicitly (based on properties of the summary) consider when selecting their preference. Towards this end, the data is processed based on the 8 users' annotations on 72 game/aspect pairs (for a total of 576 data points), and all statistical tests are performed at a 5% significance threshold. Our assumption is that the complete CL pipeline which includes both aspect detection and sentiment analysis will offer a richer and more diverse summary than AsDe alone.

Regarding users' preference of one summarization technique, results were mixed: overall, annotators had no clear preference with CL AsDe being marginally more often selected (53%). Table 4 shows the distribution of selection of CL Full split per aspect. The Table shows that the main factor for the skew of the overall preference towards CL AsDe was the graphics summaries, as the other aspects are fairly evenly preferred between the two approaches.

To further assess which factors led to the annotators' preference of one summary over the other, we conducted an analysis of variance test (ANOVA) between the preferred approach (represented as a binary choice) and other features such as the aspect. Table 5 shows the results in terms of significant differences, and verifies that there is a systematic influence between the aspect and preference. On the other hand, the game does not seem to affect users' preference of one summary or the other; this is a promising finding as the methods are supposed to be applicable to any game. There is also a clear evidence that preference was highly varying from annotator to annotator, and annotators rarely agreed with each other even in this simple pair-wise preference task.

To get a better understanding of the reasons annotators gave regarding their preference, we looked further into the statistics of the winning observations of CL AsDe vs. CL Full. When AsDe was preferred, annotators explained their preference mainly due to better coherence (63%), lower redundancy (28%), but also 'other reasons' (26%). When CL Full was preferred, annotators chose 'other reasons' (50%), and less often coherence (41%) or low redundancy (17%). This finding shows that summaries by AsDe were more coherent but annotators still preferred summaries by CL Full often for other reasons. This points to a limitation of the experimental protocol, as the interface did not provide annotators with enough options to allow them to explain their reasons for their summary preference. This was addressed in the second user study (see Section 6) with an extra option on the UI. It should be noted that better coherence was selected far more often overall (53% of instances) than lower redundancy (23%), while 'other reasons' were also chosen often (37%). Redundancy and coherence were chosen together in only 5% of instances, and thus it is evident that these two axes of evaluation are fairly independent. These findings, coupled with the statistically significant influence (via ANOVA) between preference of summarization approach and tagged coherence and redundancy, support our conclusion that both coherence and redundancy were important factors for annotators' preference.

6 SECOND USER STUDY

Based on the findings and limitations identified in the first user study, conducted a second study with more participants but fewer games, testing the best CL approaches with the novel DL Full pipeline. Due to participants' concerns on the long duration of the 72-item survey in the first experiment, we opted to use only two games to lower the time required from annotators; it is expected that fatigue would likely introduce noise to the participants' responses. Details on how the games and annotation options were chosen are detailed in Section 6.1.

6.1 Annotation Protocol

The user interface for the second user study was largely the same as in the first (see Section 5.1). Based on the first study's finding that 'other reasons' for an annotator's preference were often chosen, a fourth option was added to the UI as a tickbox stating "The summary was more focused and contained less irrelevant information." We refer to this additional option as Focus in the analysis that follows.

As noted above, to reduce the time required for the study only two games were chosen to be annotated. We chose among the games from the first user study, taking the game where CL Full had the highest preference (Tom Clancy's The Division, where CL Full was chosen 60% of the time) and the game where CL AsDe had the highest preference (Elite Dangerous, where CL AsDe was chosen 60% of the time). For each of the two games, the preferred method was chosen to present to the user, juxtaposed with the summary for the same game and aspect produced by DL Full. Therefore, the participant had to annotate 12 items, 6 aspects for Tom Clancy's the Division comparing the CL Full summary with the DL Full summary

and 6 aspects for Elite Dangerous comparing the CL AsDe summary with the DL Full summary. The rationale was to select the most successful game summaries (for both CL variants) and compare them with the novel DL pipeline. We refer to CL and DL summaries in this paper, referring to the best CL summary (CL Full or CL AsDe) as shown to the user.

As with the first user study, the order of the two options was randomized (i.e. sometimes CL summaries were shown first, sometimes second). Unlike the previous experiment, however, the order of the sentences within the same summary was also randomized; the rationale was to avoid ordering effects when the participant starts by reading an incoherent sentence first.

6.2 Participants

Fourteen participants completed this annotation task. Unlike the previous study, a snowball method for soliciting participants was followed, soliciting feedback from a broader group. Thus, this study lacks data on the demographics and gaming experience of participants, although participants were all adults and had experience in data analysis and artificial intelligence.

6.3 Results

The data collected from the experiment was a total of 168 observations. Overall CL summaries were slightly more preferred by participants (55%), although the difference is not statistically significant (Paired t-test, p-value 0.22). Interestingly, for Elite Dangerous (which was summarized by CL AsDe) the difference was more pronounced (CL AsDe preferred 60% of the time over DL Full); for Tom Clancy's The Division the two methods (CL Full and DL Full) were chosen evenly. Since only one game was tested per CL variant, it is difficult to assess whether the preference was due to the game itself or the sentiment-based selection component. Moreover, while DL Full includes sentiment-based selection, this part accounts for 2 of the 6 sentences and thus it is even more difficult to estimate the reasons for the users' preference. This ambiguity points to further refinements needed for the annotation protocol which is discussed in Section 7.

In terms of the reasons offered by participants for their choice, coherence was still most commonly chosen (62% of responses), followed closely by focus (56%). Low redundancy was chosen less often (23%), while 'other reasons' are chosen only in 14% of responses). The addition of the focus option seems to have mitigated the prevalence of 'other reasons' in the first study. Unlike the first study, however, low redundancy was often chosen in conjunction with one other reason (56% of the time) or two other reasons (36% of the time). Combined with its low overall prevalence, it is possible that low redundancy may now longer be necessary as a separate reason in the UI, although a broader user study with more games is needed to validate this hypothesis.

Pearson's Chi-squared tests were also used in order to test whether any of the above reasons is correlated to the preferred summary. Only redundancy was found to be correlated with the type of summary (p-value 0.001). This clearly indicates the importance of handling redundancy satisfyingly in any future approach.

7 DISCUSSION

This paper introduced a number of possible pipelines for identifying, grouping, and extracting the opinions of users in terms of pre-specified game facets. two small-scale user surveys examined the preference of users in the presence of different pipeline implementations. Results indicate that (a) aspect extraction is important for summarization, although deep-learning does not necessarily improve the aspect extraction process compared to a simpler clustering-based method; (b) between the clustering-based pipeline variants (CL AsDe, CL Full), there was no clear winner with respect to the summary outputs; (c) evaluators had strong and individual opinions on which variant was better; (d) sentiment-based criteria and/or confidence-based criteria for selecting sentences do not seem to perform better than the random selection performed by CL AsDe.

While the aspects chosen for this experiment were intuitive, based on typical facets of games that players and professional critics focus on, some of the resulting aspect-based summaries were less coherent than others. The choice to assign a sentence to an aspect even if its cluster only had a slim majority in keyword frequency likely introduced inconsistency. For CL aspect detection, the most significant factor for the lack of coherence was the choice of keywords. Specifically, the keyword "sound" was often found in sentences unrelated to game audio, used as a verb: e.g. "On paper this game sounds great". To a degree, such artefacts were removed in the DL aspect detection pipeline via (a) the latent sentence representation and (b) fine-tuning the model based on manual annotations on this specific corpus. However, a more sophisticated method for aspect detection seems necessary. For instance, an adaptive query expansion as followed by [29] could create a much larger set of keywords automatically, although it may overlook the nuances of game terminology. On the other hand, a Word2Vec model [30] trained on the entire corpus of steam reviews (or even larger game-related corpora such as game FAQs and fansites) could be used to derive a similarity score with specific aspects. Building a game ontology for this task or using an existing one [36, 39] could further assist in discovering more keywords or in calculating an ontology-based semantic similarity measure [42]. Finally, a completely different direction could see the discovery of topics specific to each game rather than focusing on the same pre-specified topics every time. This would be valuable as different genres have a different focus (e.g. multiplayer games focus on balance or lag, while horror games focus on the emotional response), but could make it difficult to maintain the same presentation format across games and thus confuse end-users.

Sentiment analysis was also often problematic, primarily due to the informal and idiosyncratic language that games reviews were often in. Reviews are often rife with sarcasm and negation, e.g. "Have fun spending huge amounts of hours for very little progress.". Moreover, many reviews' sentences have poor syntax and are very short or very long (e.g. "Good: + great aesthetic."). Sentiment analysis treated the sentence as a bag-of-words, exacerbating the problem. In general, sentiment analysis can not capture negation or sarcasm and handles incomplete sentences poorly. Performance would likely be improved with a more appropriate pre-trained lexicon for informal utterances on the Social Web, such as SentiStrength [46] or other

sentiment- and negation-aware approaches [22]. Alternatively, a custom classifier for sentiment analysis could be trained using text from a Steam review as input and the user's recommendation as polarity. Complementing the training set with experts' annotations could refine such a model, especially when dealing with sarcasm. Another promising alternative to SentiWordNet for sentiment analysis would be the use of an authored dictionary of opinion words [24] or game-specific adjectives annotated in terms of polarity [51].

Our findings also showed no clear winner between the two CL variants or between CL and DL summaries. These ambiguity of the findings could well be by-products of the experimental protocol followed. Findings from the first user study pointed to a missing reason for players to report, and the second study included a "focus" reason which improved the quality of the data collected but raised questions about the importance of the "low redundancy" reason. The users' reported fatigue in the first experiment led to fewer items in the second study to alleviate the burden from annotators. However, this increased the locality of the findings in the second study as it was unclear whether preferences were due to the game or the algorithm. In future studies, summaries for more games should be annotated by more participants, showing only two games to each user but randomizing which games are shown when the user starts the study. More importantly, the current experimental protocol forces participants to select one review as preferred and provide at least one reason. The forced choice between two summaries does not allow the user to provide more nuanced feedback. A four-alternative forced-choice (4-AFC) with options "A", "B", "both A and B", "neither A nor B" would allow the user to point out cases where both summaries are equally good or equally bad. The fairly even split between the two alternatives in both user studies could be due to fact that users consider some summaries shown equally bad and select randomly. On the other hand, a 4-AFC questionnaire would likely need many more participants since much of the data will be removed when no ranking is given. The need for more games, more annotation choices, and perhaps more algorithm variants (DL AsDe, for instance) point to the need for a large-scale user survey among the general gaming community, which will be performed in future work in this vein.

As discussed in Section 1 and explored on a high-level during the user study, game review summarization can be valuable both to consumers (players) and producers (game developers). However, each stakeholder has different priorities and will likely respond differently to different summary formats. The extractive summarization process was visualized as 'pure text' bullet points, which was not as engaging to either type of audience. It would be important to explore alternative visualizations for players and developers. For players, the summary could provide more structure (based on pre-specified game facets), focus more on the weights and scoring of each aspect (including visualizations such as pie-charts), show only a few polar opposites in terms of review sentences, and perhaps cross-reference these findings with other games' review summaries. For developers, on the other hand, a bottom-up topic discovery would likely be beneficial in order to identify unexpected points of contention among users. Moreover, presenting the context of the reviewers' chosen sentences would also be valuable for designers, e.g. how many reviewers agree with or echo this comment, when this comment was made and whether general sentiment has shifted since then. Such context can be important regarding the urgency of addressing certain concerns or to gauge whether patches and updates have improved reviewers' perception, not unlike Steam's use of most recent reviews.

There are many directions for future research depending on the purpose of the game review summarization. As a tool for game evaluation, primarily targeted towards players or producers, the game's context is important in order to choose which reviews or topics to highlight. Additional research in this vein would need to find topics or patterns in similar games (e.g. of the same genre, publisher, or publication date) and then to compare the current game's reviews in terms of those topics or compared to other games' reviews. User experience research would also be important to find how best to present such results, as interactive summaries where the user can zoom in and out into different games and/or different topics within games would make the summaries more intuitive and manageable. As a tool for game analysis, bottom-up probabilistic topic modelling [7] in games of the same genre could help identify design patterns [6] and players' expectations based on their repertoire [27]. As a tool for knowledge discovery, game reviews can serve as raw text or multi-modal corpora from which structured data can be automatically extracted as entities and relations [44], concept hierarchies [43, 52], or even a complete game ontology [32, 38].

8 CONCLUSION

This paper highlighted the challenges and opportunities of game review summarization via natural language processing. The paper introduced a pipeline for grouping Steam users' comments into pre-specified aspects such as visuals or performance, and studied different renderings of the final summary, exploiting positive and negative sentences based on sentiment analysis. The small-scale user survey revealed differences in how different annotators assess the reviews, highlighted possible foci of research for better game review summarization systems, and suggested a number of refinements to the process are suggested in this promising subfield of game artificial intelligence.

REFERENCES

[1] Rasim M Alguliev, Ramiz M Aliguliyev, Makrufa S Hajirahimova, and Chingiz A Mehdiyev. 2011. MCMR: Maximum coverage and minimum redundant text summarization model. *Expert Systems with Applications* 38, 12 (2011), 14514–14522.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv1409.0473 [cs, stat]* (sep 2014). https://doi.org/10.1146/annurev.neuro.26.041002.131047 arXiv:1409.0473

[4] BBC News. 2019. Gaming worth more than video and music combined. https://www.bbc.com/news/technology-46746593. Accessed 26 January 2020.

[5] Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

[6] Staffan Bjork and Jussi Holopainen. 2004. *Patterns in Game Design.* Charles River Media.

[7] David M. Blei. 2012. Probabilistic Topic Models. *Communicatiosn of the ACM* 55, 4 (2012), 77–84.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[9] Andrew P Clark, Kate L Howard, Andy T Woods, Ian S Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? Using the "EloChoice"

package to assess pairwise comparisons of perceived physical strength. *PloS one* 13, 1 (2018).

[10] Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of the Text Analysis Conference*.

[11] Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. Hillary: What Went Viral During the 2016 US Presidential Election. In *Proceedings of the International Conference on Social Informatics*. Springer International Publishing, 143–161.

[12] Sanmay Das and Mike Y. Chen. 2001. Yahoo! for Amazon: extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific finance association annual conference*.

[13] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. 9–15.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[15] Cliff Edwards. 2013. Valve Lines Up Console Partners in Challenge to Microsoft, Sony. https://www.bloomberg.com/news/articles/2013-11-04/valve-lines-up-console-partners-in-challenge-to-microsoft-sony. Accessed 26 January 2020.

[16] Eslam Elsawy, Moamen Mokhtar, and Walid Magdy. 2014. TweetMogaz v2: Identifying News Stories in Social Media. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*.

[17] Entertainment Software Association. 2018. Essential Facts About the Computer and Video Game Industry report. https://www.theesa.com/wp-content/uploads/2019/03/ESA_EssentialFacts_2018.pdf. Accessed: 5 Sep 2019.

[18] Angela Fan, David Grangier, and Michael Auli. 2017. Controllable Abstractive Summarization. In *Proceedings of the ACL Workshop on Neural Machine Translation and Generation*.

[19] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.

[20] George Giannakopoulos, George Kiomourtzis, and Vangelis Karkaletsis. 2014. Newsum:"n-gram graph"-based summarization in the real world. In *Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding*. IGI Global, 205–230.

[21] George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 270–274.

[22] Maria Giatsoglou, Manolis G Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch. Chatzisavvas Chatzisavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications* 69 (2017), 214–224.

[23] Hongyu Han, Yongshi Zhang, Jianpei Zhang, Jing Yang, and Xiaomei Zou. 2018. Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias. *PLOS ONE* 13 (2018).

[24] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 168–177.

[25] Ya-Han Hu, Yen-Liang Chen, and Hui-Ling Chou. 2017. Opinion mining from online hotel reviews – A text summarization approach. *Information Processing & Management* 53, 2 (2017), 436–449.

[26] San-Yih Hwang, Chia-Yu Lai, Jia-Jhe Jiang, and Shanlin Chang. 2014. The identification of Noteworthy Hotel Reviews for Hotel Management. *Pacific Asia Journal of the Association for Information Systems* 6 (2014).

[27] Jesper Juul. 2005. *Half Real. Videogames between Real Rules and Fictional Worlds*. MIT Press.

[28] Antonios Liapis, Georgios N. Yannakakis, Mark J. Nelson, Mike Preuss, and Rafael Bidarra. 2019. Orchestrating Game Generation. *IEEE Transactions on Games* 11, 1 (2019), 48–68.

[29] Walid Magdy and Tamer Elsayed. 2016. Unsupervised adaptive microblog filtering for broad dynamic topics. *Information Processing & Management* 52, 4 (2016), 513–528.

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[31] Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Feature Specific Sentiment Analysis for Product Reviews. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 475–487.

[32] Roberto Navigli and Paola Velardi. 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics* 30, 2 (2004), 151–179.

[33] Chikashi Nobata and Satoshi Sekine. 2004. CRL/NYU Summarization System at DUC-2004. In *Document Understanding Workshop 2004*.

[34] Karolina Owczarzak, John M Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. Association for Computational Linguistics, 1–9.

[35] George Panagiotopoulos, George Giannakopoulos, and Antonios Liapis. 2019. A Study on Video Game Review Summarization. In *Proceedings of the MultiLing Workshop*.

[36] Janne Parkkila, Filip Radulovic, Daniel Garijo, María Poveda-Villalón, Jouni Ikonen, Jari Porras, and Asuncion Gomez-Perez. 2016. An ontology for videogame interoperability. *Multimedia Tools and Applications* 76 (2016).

[37] M.F. Porter. 2006. An algoritm for suffix stripping. *Program* 14 (2006), 130–137.

[38] Ligaj Pradhan, Chengcui Zhang, and Steven Bethard. 2016. Towards extracting coherent user concerns and their hierarchical organization from user reviews. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*. IEEE, 582–590.

[39] Owen Sacco, Antonios Liapis, and Georgios N. Yannakakis. 2017. Game Character Ontology (GCO): A Vocabulary for Extracting and Describing Game Character Information from Web Content. In *Proceedings of the International Conference on Semantic Systems*.

[40] Hassan Saif, Miriam Fernandez, and Harith Alani. 2015. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management* 52 (2015).

[41] Gerard Salton and C.S. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation*. 29, 4 (1973), 351–372.

[42] David Sánchez, Montserrat Batet, David Isern, and Aida Valls. 2012. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications* 39 (2012).

[43] Mark Sanderson and W. Bruce Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

[44] Min Song, Won Chul Kim, Dahee Lee, Go Eun Heo, and Keun Young Kang. 2015. PKDE4J: Entity and relation extraction for public knowledge discovery. *Journal of Biomedical Informatics* 57 (2015), 320 – 332.

[45] Pero Subasic and Alison Huettner. 2001. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems* 2 (2001), 483 – 496.

[46] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology* 63, 1 (2012), 163–173.

[47] Peter D. Turney. 2002. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[49] Janyce Wiebe. 2000. Learning Subjective Adjectives from Corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. 735–740.

[50] Kevin Yauris and Masayu Leylia Khodra. 2017. Aspect-based summarization for game review using double propagation. In *Proceedings of the International Conference on Advanced Informatics, Concepts, Theory, and Applications*.

[51] José P. Zagal, Noriko Tomuro, and Andriy Shepitsen. 2012. Natural Language Processing in Game Studies Research: An Overview. *Simulation & Gaming* 43, 3 (2012), 356–373.

[52] Elias Zavitsanos, Georgios Paliouras, George A Vouros, and Sergios Petridis. 2010. Learning subsumption hierarchies of ontology concepts from texts. *Web Intelligence and Agent Systems: An International Journal* 8, 1 (2010), 37–51.

[53] Lili Zhao and Chunping Li. 2009. Ontology Based Opinion Mining for Movie Reviews. In *Knowledge Science, Engineering and Management*. Springer Berlin Heidelberg, 204–214.

[54] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie Review Mining and Summarization. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 43–50.

[55] Zhen Zuo. 2018. *Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier*. Technical Report. University of Illinois at Urbana–Champaign.