Creating a Sentiment Lexicon with Game-Specific Words for Analyzing NPC Dialogue in The Elder Scrolls V: Skyrim

Thérèse Bergsma, Judith van Stegeren, Mariët Theune
Human Media Interaction, University of Twente
Enschede, The Netherlands
tslbergsma@gmail.com, j.e.vanstegeren@utwente.nl, m.theune@utwente.nl

Abstract
A weak point of rule-based sentiment analysis systems is that the underlying sentiment lexicons are often not adapted to the domain of the text we want to analyze. We created a game-specific sentiment lexicon for video game Skyrim based on the E-ANEW word list and a dataset of Skyrim’s in-game documents. We calculated sentiment ratings for NPC dialogue using both our lexicon and E-ANEW and compared the resulting sentiment ratings to those of human raters. Both lexicons perform comparably well on our evaluation dialogues, but the game-specific extension performs slightly better on the dominance dimension for dialogue segments and the arousal dimension for full dialogues. To our knowledge, this is the first time that a sentiment analysis lexicon has been adapted to the video game domain.

Keywords: sentiment analysis, sentiment lexicon, ANEW, video games, dialogue, lore, Skyrim, E-ANEW

1. Introduction
Sentiment analysis is a subfield of NLP that tries to assign sentiment ratings to texts. A drawback of rule-based sentiment analysis methods is that sentiment lexicons, the lists that link specific words to sentiment values, are often not adapted to the specific domain of the text. In this research we investigate whether we can adapt an existing sentiment analysis lexicon to a sentiment lexicon for the video game domain.

Sentiment analysis for video games can be used as a stepping stone to achieve affective language recognition for game texts (such as NPC dialogue). If we can automatically distinguish positive and negative polarity in game texts, we might be able to extend this to multi-dimensional emotion recognition. This can help us to automatically assess opinion, emotion and personality of individual characters in video games. Additionally, these metrics could be a first step towards quantitative analysis for games as a whole, based on their texts. Holistic text-based metrics could be used to compare multiple games in the same genre, or find similarities between games across multiple domains. Finally, we expect that improved NLP methods for game texts can also inform natural language generation for games.

In this research we use an extended version of the Affective Norms for English Words (ANEW) list as the basis for the sentiment analysis lexicons. The Affective Norms for English Words (ANEW) list was published by Bradley and Lang (1999). It contains 1034 English words with a normative emotional rating. Bradley and Lang (1999) followed for their normative emotional rating the work of Osgood et al. (1957), who found that variance in emotional assessments can be captured in three major dimensions: valence (or polarity), arousal and dominance. These dimensions are often named the PAD dimensions in literature. See Figure 1 for an overview.

The 1034 ANEW words received their ratings from introductory psychology class students who participated as part of a course requirement. Each student rated words for all three dimensions. Warriner et al. (2013) added more words to the original ANEW word list and gathered ratings for each word through Amazon Mechanical Turk (MTurk) which resulted in an extended ANEW (E-ANEW) word list of 13,915 rated words. Almost all previously ANEW rated words by Bradley and Lang (1999) were also included and received a new ANEW rating from the participants of MTurk. Here participants rated words for just one dimension, so a participant focused on only providing ratings for either valence, arousal or dominance. Each word received between 14 and 109 ratings for a dimension.

One possible use for ANEW rated words is predicting sentence sentiment by looking at the sentiment of the words that are in the sentence (Gökçay et al., 2012). However, ANEW consists only of words that are included in an English dictionary. Consequently, sentiment analysis for texts that contain many technical words, made-up words or abbreviations might not perform as well for texts that contain only common English words.

One category of texts that might contain made-up words is texts from video games. Especially role-playing games (RPGs), for example those with a fantasy or science fiction setting, often use invented words to enrich the story and to help with world building. One example of such an RPG is The Elder Scrolls V: Skyrim (Bethesda Game Studios 2011), an action role-playing game (action RPG) and the fifth installment in The Elder Scrolls (TES) game series. Skyrim is widely popular. It sold 30 million copies in the late 2011. However, only human raters rated the Skyrim speeches, this is a limitation of this study. In the next section we explain how we created a game-specific sentiment lexicon for Skyrim.

Figure 1: PAD dimensions overview
first five years of its release in 2011, and on average, players spend 150 hours in *Skyrim* (Suellentrop, 2016). The game has an open-world setting: players have the freedom to travel to any place in the game at any time, and quests can be completed in any order or even ignored completely. The world in *Skyrim* is populated by 1087 non-playable characters (NPCs). Together they have over 60,000 lines of dialogue (Senior, 2011). In *Skyrim*, all NPCs have their own daily routines: they go to bed at certain times, do chores and run errands, and talk to other NPCs. NPCs also function as quest-givers in the game. They might ask the player to do something for them, such as retrieving a family heirloom from a cave filled with necromancers, or accompanying them during their travels on a dangerous road.

Although NPCs speak English, some of the words they use cannot be found in an English dictionary. Many words are made-up and are part of *The Elder Scrolls*’ lore that the developers created to enrich the game world for the players. This research focuses on sentiment analysis of dialogue of NPCs. We want to find out whether existing (such as ANEW or E-ANEW), or extended (such as lexicons adapted to a game, a game series or genre) English sentiment analysis lexicons are suitable to compute sentiment ratings for NPC dialogue in a fantasy setting. Specifically, we consider a lexicon to be suitable if it produces sentiment ratings that are comparable to those given by human raters. We compare human sentiment ratings of *Skyrim* dialogue with ratings computed using the E-ANEW lexicon and a *Skyrim*-specific extension of E-ANEW.

2. Related work

2.1. Domain-specific sentiment lexicons

Studying and comparing the performance of a genre specific lexicon is especially popular for social media text. Nielsen (2011) examined the performance of sentiment analysis for Twitter by using the ANEW list and comparing this with sentiment analysis that uses a new word list that was specifically constructed for the language that is often seen on micro-blogs, such as Internet slang acronyms (e.g. LOL and WTF) and obscene words.

Similarly, Hutto and Gilbert (2014) created VADER (Valence Aware Dictionary for sEntiment Reasoning), a rule-based model for sentiment analysis on social media texts. In their lexicon they included emoticons (e.g. :-)) slang with sentimental value (e.g. nah and meh) and acronyms. This resulted in a lexicon with 7,500 words. Each entry was rated by ten human raters. Aside from using a self-made lexicon, VADER also takes into account how syntax and punctuation can influence the perceived sentiment of a text.

2.2. Computing sentiment score from word ratings

In our research we computed sentiment scores by identifying words and their PAD values, summing these values for each PAD dimension and dividing them by the number of words with a PAD value in the sentence. Taking the mean to compute a sentiment score has previously been done by (Guerrini et al., 2012) to measure the message impact of Google AdWords and (Staiano and Guerrini, 2014) to perform sentiment analysis on crowd-annotated news. Nielsen (2011) also tried variations on the computation by not normalizing words, by normalizing the sum by looking only at words with a non-zero valence value (a technique that diminishes returns of values with a neutral meaning (Guerrini et al., 2012)), by only taking into account the words with the most extreme valence values, and by changing the valence values to $+1$, $0$ and $-1$.

2.3. NLP on game texts

Most research on games and NLP deals with generating textual content for games; there is only limited work on analyzing game texts. Louis and Sutton (2018) created language models of character and action descriptions in a role-playing game (RPG) with the goal of inferring the latent ties between actions and characters. Urbanek et al. (2019) crowdsourced a collection of player interactions in a fantasy text adventure, and used it to train generative and retrieval models to predict action, emote, and dialogue sequences. Kerr and Szafrań (2009) present a machine learning approach for classifying the level of sophistication of dialogue lines (trained on movie dialogues), which they tested on a manually annotated collection of dialogue lines from the game Neverwinter Nights.

Existing work on sentiment analysis in a game context is generally aimed at the analysis of texts produced by players of video games rather than NPCs. Fraser et al. (2018) perform sentiment analysis on utterances of users talking to a character based on an NPC from *The Elder Scrolls V: Skyrim*. Others apply sentiment analysis to game-related tweets (Sarratt et al., 2014), game reviews (Borgholt et al., 2015), Panagiotopoulou et al., 2019), and chat messages on a video game streaming platform (Barbieri et al., 2017). Whether these texts are about games or used inside games, they are likely to include game-specific words, and thus all these applications could potentially benefit from using a game-specific sentiment lexicon. We are not aware of any previous work on using a game-specific sentiment lexicon for sentiment analysis.

3. Data

For this research we used the following language resources:

1. lore from *The Elder Scrolls*: a dataset of in-game books, letters and notes from games in *The Elder Scrolls* series, scraped from a fan-website
2. a plaintext file with *Skyrim* dialogue lines
3. the extended ANEW word list from Warriner et al. (2013) with 13,915 rated words
4. an English lexicon

3.1. *The Elder Scrolls*’ lore

The first dataset contains the text of all in-game documents from games in *The Elder Scrolls*. The collection consists of 4,890 books, letters and notes. The texts were scraped from *The Imperial Library* website [a website that collects in-game documents from The Elder Scrolls.](https://www.imperial-library.info/books/all/by-category)
A is for Atronach.
B is for Bungler’s Bane.
C is for Comberry.

Figure 2: Example book *ABCs for Barbarians* from the Imperial Library dataset

| Property      | Value                                       |
|---------------|---------------------------------------------|
| Conversation  | FormID: 000AF489                            |
| Quest ID      | DialogueSolitude                            |
| Quest branch  | DialogueSolitudeFalkBranch3                 |
| Quest topic   | DialogueSolitudeFalkBranch3Topic            |
| Subtype       | CUST                                        |
| Id            | 0                                           |
| Text          | Of course he does. What sort of a question is that? |
| Notes         | Annoyed                                     |

Figure 3: An example entry from the Thuum.org dialogue dataset

### 3.2. *Skyrim* dialogue file

The second resource is a file with in-game dialogue from *Skyrim*. The dataset was acquired from Thuum.org[3](https://www.thuum.org/), a fan website that is dedicated to documenting *Skyrim*’s dragon language. Thuum provides a text file[4](https://www.thuum.org/library/Dialogue), which contains more than 34,000 dialogue entries from the base game.

Each entry consists of a conversation identifier, a quest identifier, a quest branch identifier, the topic of the quest, a dialogue subtype, an identifier for the entry, 1-3 lines of dialogue, and optional direction notes for voice actors.

### 3.3. Extended ANEW word list

The extended ANEW word list (Warriner et al., 2013) contains 13,915 words and their ratings. Ratings range from 1 (unpleasant, calm, controlled) to 9 (pleasant, excited, in control), with 5 signifying a neutral rating.

Each entry contains 65 values, listing the mean valence, arousal and dominance ratings given by participants, standard deviation (SD), the number of ratings each word received during the crowd-sourcing experiment, as well as those values for different demographic groups of participants: all participants, females, males, higher educated, lower educated, old people and young people.

Only a few of the 65 values from the E-ANEW dataset are necessary for this research: the word itself, the mean rating and standard deviation from all participants for valence, arousal and dominance.

### 3.4. English words lexicon

We used a public lexicon of 466,549 English words. It was retrieved from a GitHub repository[5](https://github.com/dwyl/english-words) and is based on the dataset from Project Gutenberg[6](https://www.gutenberg.org/) which contains mostly public domain books with more than 57,000 items, and the Moby Project[7](https://github.com/dwyl/english-words), which consists of more than 177,000 words and their pronunciation.

### 4. Game-specific sentiment lexicon

In order to investigate whether sentiment analysis for NPC dialogue improves when we use a game-specific sentiment analysis lexicon (E-ANEW-TES), we want to compare human sentiment ratings for NPC dialogue with sentiment ratings calculated using E-ANEW, and a *Skyrim*-specific extension of E-ANEW. In this section we explain how we created the latter.

#### 4.1. Creating a game-specific word list

We tokenized and stemmed the text of the dataset with TES lore, filtered out unique words and ordered these by frequency. All words that occurred in the English lexicon were removed. We manually removed English words that were not part of our English lexicon, such as the word *false*. This left us with 13,831 words considered to be unique to *The Elder Scrolls*, most of which are nouns and names.

We created a custom stemmer, i.e. a mapping from word variations to their stems, for *Skyrim*-specific words with a naive unsupervised approach: we checked whether words in our word list (i.e. stems) also occurred with additional letters appended (i.e. variations). This way we could detect both regular plural nouns, such as *nirnroot* (singular) → *nirnroots* (plural) and related adjectives, such as *khajiit* (singular) → *khajiiti* (adjective). Filtering out word variations led to a more representative frequency list of words. Finally, we removed all words that occurred less than 10 times. The final list contained 965 word stems and an additional 119 word variations.

For an example of word list entries, see Figure[7](https://github.com/dwyl/english-words)

#### 4.2. Calculating sentiment ratings with word2vec

The next step lies in providing each word in our game-specific word list with a value for valence, arousal and dominance, so that we can use them for calculating a sentiment rating for a line of NPC dialogue. From here on, we call the *Skyrim* lexicon with PAD values *E-ANEW-TES*, where TES stands for *The Elder Scrolls*.

We use the ratings of the E-ANEW words to extrapolate a sentiment rating for game-specific words. Specifically, we average the ratings of the three most similar E-ANEW words to calculate the PAD values for a game-specific word. We use word2vec (Mikolov et al., 2013) to find the most similar E-ANEW words for each game-specific word. Word2vec turns words into vectors that represent the context of those words. We can use those vectors to find similar words in a dataset: words that are used in similar contexts
We calculated the PAD values for each word in our validation set. All The Elder Scrolls games to date have focused on the continent of Tamriel.

We calculate the PAD values for a game-specific word by applying word2vec on the lore dataset to obtain word vectors. We applied word2vec on the lore dataset to obtain word vectors for each of the game-specific words. We use a context window size of eight (i.e. for each word, the algorithm takes the eight words before and after that word into account) to find words that are more semantically close than only topically related (Jurafsky and Martin, 2008).

We calculate the PAD values for a game-specific word by averaging the PAD values of the three E-ANEW words that have the highest probability of occurring in the context of the game-specific word. If we use more than three words to calculate the PAD values, averaging the values means we might end up with mostly neutral PAD values for game-specific words. On the other hand, using three words for calculation mitigates the possibility of word2vec mistakes, and prevents that one ANEW word determines the PAD values too much.

4.3 Validating the word2vec model

Fifteen words from the E-ANEW list that also occur in the Skyrim lore (e.g. sword, inn, werewolf in contrast to phone, satellite, streetcar) were randomly chosen as validation set to determine whether the model provides good E-ANEW-TES ratings with the TES lore as training data. We filtered out English words that have a specific meaning in the context of Skyrim, such as shout (‘dragon shouts’ are a form of magic in Skyrim), cat (Skyrim contains humanoid cats) and empire (the Empire is a political entity and denotes a faction name in The Elder Scrolls).

We calculated the PAD values for each word in our validation set, so we could gauge whether the PAD values are within acceptable limits. A PAD value for E-ANEW-TES is considered satisfactory if the value stays within the boundaries of the standard deviation (SD) of the E-ANEW rating of that same word. For example, the E-ANEW valence rating for sword is 5.27 and the SD is 1.58, meaning that an E-ANEW-TES rating for sword will only be considered satisfactory if it is somewhere between 3.69 and 6.85, which is with an E-ANEW-TES rating of 4.43. We retrained the word2vec model until the results provided all fifteen words with a satisfactory rating. Another fifteen random words were selected and tested which also stayed within the SD of the E-ANEW valence ratings with the TES lore as training data. We filtered out English words that have a specific meaning in the context of Skyrim, such as shout (‘dragon shouts’ are a form of magic in Skyrim), cat (Skyrim contains humanoid cats) and empire (the Empire is a political entity and denotes a faction name in The Elder Scrolls).

We calculated the PAD values for each word in our validation set, so we could gauge whether the PAD values are within acceptable limits. A PAD value for E-ANEW-TES is considered satisfactory if the value stays within the boundaries of the standard deviation (SD) of the E-ANEW rating of that same word. For example, the E-ANEW valence rating for sword is 5.27 and the SD is 1.58, meaning that an E-ANEW-TES rating for sword will only be considered satisfactory if it is somewhere between 3.69 and 6.85, which is with an E-ANEW-TES rating of 4.43. We retrained the word2vec model until the results provided all fifteen words with a satisfactory rating. Another fifteen random words were selected and tested which also stayed within the SD of the E-ANEW valence ratings with the TES lore as training data. We filtered out English words that have a specific meaning in the context of Skyrim, such as shout (‘dragon shouts’ are a form of magic in Skyrim), cat (Skyrim contains humanoid cats) and empire (the Empire is a political entity and denotes a faction name in The Elder Scrolls).

5. Sentiment analysis on game text

5.1 Calculating game text ratings

To obtain a sentiment rating for a game text, e.g. a piece of NPC dialogue, we follow the calculation method of Nielsen (2003). We repeated this procedure for each of the 965 words of our game-specific word list. The words with their respective PAD values, together with the words from E-ANEW, make up our game-specific sentiment lexicon: E-ANEW-TES. We repeated this procedure for each of the 965 words of our game-specific word list. The words with their respective PAD values, together with the words from E-ANEW, make up our game-specific sentiment lexicon: E-ANEW-TES.
Figure 6: This table shows the words that, according to our word2vec model, have the highest probability of occurring in the context of the word *septim* in *The Elder Scrolls*’ lore. The first three words are names of NPCs related to the Septim dynasty. *Divines* is a reference to the Nine Divines, another name for the pantheon of the Empire, of which Tiber Septim is a member.

| Word      | Probability       |
|-----------|-------------------|
| reman     | 0.39224135875701904 |
| katariah  | 0.3528633713722229  |
| cassynder | 0.3425233641052246  |
| throne    | 0.328492671251297   |
| empire    | 0.321741045213318   |
| divines   | 0.32171204686164856 |

Figure 7: To calculate the valence, arousal and dominance values for the word *septim*, we average the mean PAD value of the three related E-ANEW words *throne*, *empire* and *emperor*.

```
| Word | Valence | Arousal | Dominance |
|------|---------|---------|-----------|
| throne | 5.45 | 5.22 | 6.19 |
| empire | 5.36 | 4.59 | 5.95 |
| emperor | 4.68 | 4.25 | 5.32 |
| septim | 5.163 | 4.686 | 5.82 |
```

Figure 8: A dialogue entry from the *Skyrim* dialogue file that contains the word *Septim*. Each bold word has an E-ANEW rating and the underlined word has an E-ANEW-TES rating. The table underneath shows how the sentiment rating for valence is calculated for each lexicon.

```
| Word | Valence |
|------|---------|
| year | 5.15 |
| emperor | 4.68 |
| be | 6.18 |
| murder | 1.48 |
| temple | 5.3 |
| one | 6.09 |
| imperial | 4.50 |
| city | 6.12 |
| 3E | 4.50 |
| septim | 5.16 |
```

```
| Mean valence | 5.01 |
| Mean valence | 4.97 |
```

“**In the year 3E 41, Emperor Pelagius Septim** was murdered in the **Temple of the One** in the **Imperial City**. Cut down by a **Dark Brotherhood assassin**.”

[2011]: we search the game text for words from our sentiment lexicon and for each sentiment dimension (valence, arousal and dominance) separately, we average the values of these words to obtain a sentiment rating for the game text as a whole.

See Figure 8 for an example of a dialogue snippet for which we want to obtain a sentiment rating, and a table that shows how the sentiment rating is calculated for valence.

5.2. Selecting evaluation dialogue

To be able to compare the performance of E-ANEW and E-ANEW-TES, we need to apply both sentiment lexicons to dialogue snippets that contain game-specific words. We also need to keep in mind that we want humans to rate the dialogue snippets as well, so we can compare the sentiment lexicon performance to a gold standard of human ratings. There were various approaches we could take for selecting evaluation dialogue. We could source dialogue snippets from *Skyrim*’s main questline, which contains roughly 35,000 words of dialogue. However, the main questline contains only one *Skyrim*-specific word for every 33 words of text. This is especially problematic for when we want to compare E-ANEW and E-ANEW-TES performance with human ratings: this would mean that our human raters have to rate many dialogue snippets before we have collected enough information to properly evaluate the performance. We could pick a different quest, but this has another clear drawback: a quest tends to repeat the same *Skyrim*-specific words instead of touching upon a broad variety. To illustrate:

```
even *Skyrim*’s main quest consists of more than 1000 *Skyrim*-specific words, but only 76 of those are unique. If only a small part of a quest is selected, then the variation decreases even further.
```

Another approach would be to single out the sentences with a higher *Skyrim*-specific word occurrence. However, providing human raters with one sentence in isolation could make it more difficult for them to understand the context in which the sentence is said, which might negatively influence the quality of human ratings.

We decided to evaluate the sentiment lexicons on multiple independent dialogues, i.e. dialogues that do not belong to the same quest or the same NPCs. The dialogues were grouped by their identification number, see Section 3.2., and filtered on length. All dialogues should consist of four dialogue segments (entries from the *Skyrim* dialogue file that consist of 1-3 sentences each). Additionally, evaluation dialogues should contain at least three unique game-specific words. We selected five dialogues from the remaining subset for our evaluation. For an example dialogue, see Figure 10. All evaluation dialogues can be found in Figure 9. In the rest of this paper, we will refer to each segment or complete dialogue with the corresponding id from that table.

In addition to rating each dialogue as a whole, we also rated each dialogue segment separately.

6. Evaluation

To evaluate the performance of both the E-ANEW and E-ANEW-TES sentiment lexicons, we compare their sentiment ratings for the evaluation dialogues with those of human raters.
6.1. Collecting human ratings

Since [Warriner et al. (2013)] collected at least 14 ratings per word for the E-ANEW lexicon, we aimed for at least 14 participants. In order to collect representative ratings for snippets with game-specific words, we searched specifically for participants that were familiar with *Skyrim*.

We collected sentiment ratings from a group of 15 participants via a digital questionnaire. Participants were asked to rate all dialogue segments and all complete dialogues on all three PAD dimensions.

To fulfill the familiarity requirement, the instructions began with a list of twenty *Skyrim*-specific terms occurring in the evaluation dialogues. This list contained both terms that are part of the 965 *Skyrim*-specific words used for this research (e.g. *Hammerfell*, *Ulfric*) and words that are unique for *The Elder Scrolls* series but that are not part of that list (e.g. *the Blades, Summerset Isle*). Participants were encouraged to search the Internet for more information if they did not know a particular term. If a participant was familiar with the meaning of a term, they could check the box in front of it. Checking off all game-specific terms was a prerequisite for continuing with the rating.

We provided participants with a short description of the context of each dialogue. For an example, see Figure 6. All participants rated all five dialogues. For each dialogue they gave fifteen ratings; they rated each dialogue segment separately, of which there are four, and the dialogue as a whole. Some left rather than endure the growing hatred from the local Nords.

Only the *College* really remains.

We collected sentiment ratings from a group of 15 participants via a digital questionnaire. Participants were asked to rate all dialogue segments and all complete dialogues on all three PAD dimensions.

To fulfill the familiarity requirement, the instructions began with a list of twenty *Skyrim*-specific terms occurring in the evaluation dialogues. This list contained both terms that are part of the 965 *Skyrim*-specific words used for this research (e.g. *Hammerfell*, *Ulfric*) and words that are unique for *The Elder Scrolls* series but that are not part of that list (e.g. *the Blades, Summerset Isle*). Participants were encouraged to search the Internet for more information if they did not know a particular term. If a participant was familiar with the meaning of a term, they could check the box in front of it. Checking off all game-specific terms was a prerequisite for continuing with the rating.

We provided participants with a short description of the context of each dialogue. For an example, see Figure 6. All participants rated all five dialogues. For each dialogue they gave fifteen ratings; they rated each dialogue segment separately, of which there are four, and the dialogue as a whole. Some left rather than endure the growing hatred from the local Nords.

Only the *College* really remains.
We calculated the correlation between the human ratings and ANEW and the human ratings and E-ANEW-TES are comparable. For segments, the results are the same. For the complete dialogues, there is a satisfactory rating (out of 5 total) for respectively valence, arousal and dominance. For E-ANEW, 11, 17 and 20 dialogue segments have a satisfactory rating (out of 20 total) and 4, 5 and 5 complete dialogues have a satisfactory rating (out of 5 total) for valence, arousal and dominance. For E-ANEW-TES the results are comparable. For segments, the results are the same. For the complete dialogues, there is one less dialogue with a satisfactory rating for valence. We underestimated the Thalmor, and they destroyed us.

Figure 10: A dialogue from our evaluation set, together with the context that was presented to participants in our experiment. Aldmeri, Valenwood, Thalmor, Tamriel and Elsweyr are Skyrim-specific words with an E-ANEW-TES rating.

### 6.2. Comparing lexicon ratings with human ratings

To compare the performance of the two lexicons, we consider both the number of satisfactory ratings and the correlation between human sentiment ratings and the ratings computed with the lexicons. A calculated E-ANEW or E-ANEW-TES rating is considered satisfactory if it stays within the boundaries of the standard deviation (SD) of the human rating for the same dialogue or dialogue segment. For an example, see Section 4.3.

Figure 12 shows the number of satisfactory ratings for both lexicons. For E-ANEW, 11, 17 and 20 dialogue segments have a satisfactory rating (out of 20 total) and 4, 5 and 5 complete dialogues have a satisfactory rating (out of 5 total) for valence, arousal and dominance. For E-ANEW-TES the results are comparable. For segments, the results are the same. For the complete dialogues, there is one less dialogue with a satisfactory rating for valence. We calculated the correlation between the human ratings and the sentiment ratings computed with each lexicon. Since it is not possible with only 15 participants to determine whether the human ratings are normally distributed, we calculated the correlation with a metric for parametric variables (Pearson correlation) and one for non-parametric variables (Spearman correlation). However, in both cases the results are comparable, with a slightly better performance by our game-specific lexicon E-ANEW-TES.

The correlation scores between the human ratings and E-ANEW and the human ratings and E-ANEW-TES are presented in Figure 13. For each correlation, we also report the significance as a 2-tailed p-value. The p-value was calculated using SciPy’s built-in significance test\(^{11}\) for the Pearson correlation and the Spearman correlation.

### 7. Discussion

If we look at the results in Figure 12 and Figure 13, we cannot conclusively state that our game-specific lexicon performs significantly better than E-ANEW. If we look only at satisfactory ratings, E-ANEW-TES performs slightly worse than E-ANEW, since it has one less dialogue segment with a satisfactory rating for valence. When we look at Pearson correlation scores between human ratings and the two lexicons, E-ANEW performs better for the valence dimension, and E-ANEW-TES performs better on arousal and dominance, independent of the length (e.g. segments or complete dialogues) of the evaluation text.

If we consider the Spearman correlation, E-ANEW-TES performs slightly better than E-ANEW. It performs better than E-ANEW on dominance for segments and on arousal for whole dialogues. E-ANEW performs better on valence for segments. For all other sentiment dimensions and text types, both lexicons perform equally well. However, except for arousal, most correlation scores are not significant. This is probably due to the small amount of participants in our evaluation experiment.

When we calculate a sentiment rating for a dialogue segment or a complete dialogue, we take only a few game-specific words into account. On average, this was one word per dialogue segment (with a maximum of three) and five words per complete dialogue (with a maximum of eight words). This low density of recognized game-specific words makes it hard for E-ANEW-TES to perform significantly better than E-ANEW. However, analyzing texts with a large number of recognized words (both E-ANEW words and game-specific words) can also become problematic. Since we average the values of all recognized words, recognizing more words in a text will inevitably lead to the ratings that approach a neutral rating of 5. Similarly, longer texts might also contain more words from our lexicon, which also invariably leads to more neutral ratings and decreased performance.

We selected our evaluation dialogues with the evaluation process in mind: dialogues were selected for containing at least four game-specific words, with at least three of those words unique to the dialogue. However, most dialogues from the game will not satisfy these constraints. This means that when choosing a Skyrim dialogue at random, the performance might be worse than in the ideal situation used for this research.

### 8. Improvements

There are multiple possibilities for improving both our game-specific lexicon E-ANEW-TES and sentiment analysis for games in general.

A possible improvement is creating a game-specific lexicon that also includes n-grams, i.e. game-specific terms that consist of multiple words. For example, the bigrams Oblivion Crisis and Red Year from [d2] would receive four E-ANEW ratings, since oblivion, crisis, red and year all occur in its lexicon. However, the bigrams have their own particular meaning in *The Elder Scrolls*. Additionally, words that are names in *The Elder Scrolls* but also appear in our English

\(^{11}\) For more information, see the documentation of SciPy 1.4.1. [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html)
|                      | Dialogue segments | Complete Dialogues |
|----------------------|-------------------|-------------------|
|                      | Valence | Arousal | Dominance | Valence | Arousal | Dominance |
| Highest rating       | 6       | 7.133   | 5.533     | 5.667   | 6.133   | 5.667     |
| Lowest rating        | 2.667   | 3.533   | 3.467     | 3.267   | 4       | 3.533     |
| Mean rating          | 4.257   | 4.783   | 4.390     | 4.187   | 5.08    | 4.387     |
| Mean SD              | 1.414   | 1.920   | 2.301     | 1.392   | 1.826   | 2.020     |

Figure 11: Human ratings for dialogue segments and complete dialogues obtained from 15 participants. Sentiment ratings range from 1 (unpleasant, calm, controlled) to 9 (pleasant, excited, in control).

|                      | Number of satisfactory ratings |
|----------------------|-------------------------------|
|                      | E-ANEW | E-ANEW-TES |
|                      | Segments |          |          |
| Valence              | 11      | 11        |
| Arousal              | 17      | 17        |
| Dominance            | 20      | 20        |
|                      | Dialogues |          |          |
| Valence              | 4       | 3         |
| Arousal              | 4       | 4         |
| Dominance            | 5       | 5         |

Figure 12: Number of segments and dialogues with a satisfactory sentiment rating, as calculated using the E-ANEW and E-ANEW-TES lexicons. The results of the two lexicons are comparable.

Word list (e.g. *Empire* and *nord*) often have a different meaning in *Skyrim*. This is not reflected in the current E-ANEW-TES word list. Additionally, future work should focus on larger-scale evaluation with more participants, more evaluation dialogues and texts of different lengths. It would also be interesting to apply the same method on in-game text from other games. Finally, instead of simply averaging the sentiment ratings of the lexicon words, we could apply more sophisticated methods for calculating the sentiment rating for a text.

9. Conclusion

The inclusion of game-specific or genre-specific words in a sentiment lexicon seems a suitable approach for improving sentiment analysis for games. However, text for which we want to calculate a sentiment rating should have a high density of game-specific words before using a game-specific lexicon makes a noticeable difference.

E-ANEW-TES, the E-ANEW extension that also includes game-specific words performed better on complete *Skyrim* dialogues than E-ANEW. However, the performance difference between E-ANEW-TES and E-ANEW is very small. In most cases, the performance of E-ANEW-TES was the same as that of E-ANEW.

The code, datasets and results of this research are available on Github: [https://github.com/jd7h/sentiment-lexicon-skyrim](https://github.com/jd7h/sentiment-lexicon-skyrim)

10. Acknowledgments

This research is partially supported by the Netherlands Organisation for Scientific Research (NWO) via the DATA2GAME project (project number 055.16.114).

11. References

Barbieri, F., Espinosa-Anke, L., Ballesteros, M., Soler-Company, J., and Saggion, H. (2017). Towards the understanding of gaming audiences by modeling Twitch emotes. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 11–20. Association for Computational Linguistics.

Bethesda Game Studios. (2011). *The Elder Scrolls V: Skyrim*. Game [PC]. Bethesda Softworks, Rockville, Maryland, US.

Borgholt, L., Simonsen, P., and Hovy, D. (2015). The rating game: Sentiment rating reproducibility from text. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2527–2532. Association for Computational Linguistics.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, Center for Research in Psychophysiology, University of Florida.

Fraser, J., Papaioannou, I., and Lemon, O. (2018). Spoken conversational AI in video games: Emotional dialogue management increases user engagement. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA ’18, page 179–184. Association for Computing Machinery.

Gökçay, D., İşbilir, E., and Yıldırım, G. (2012). Predicting the sentiment in sentences based on words: An exploratory study on ANEW and ANET. In 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), pages 715–718. IEEE.

Guerini, M., Strapparava, C., and Stock, O. (2012). Ecological evaluation of persuasive messages using Google AdWords. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 988–996. Association for Computational Linguistics.

Hutto, C. J. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI Conference on Weblogs and Social Media*, pages 216–225.

Jurafsky, D. and Martin, J. (2008). *Speech and Language Processing*. Pearson Education (US).

Kerr, C. and Szafron, D. (2009). Supporting dialogue generation for story-based games. In *Proceedings of the Fifth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE ’09, page 154–160. AAAI Press.

Louis, A. and Sutton, C. (2018). Deep Dungeons and Drag-
|          | E-ANEW       | E-ANEW-TES   |
|----------|--------------|--------------|
|          | Pearson      | p-value      | Pearson      | p-value      |
|          | correlation  |              | correlation  |              |
| Segments |              |              |              |              |
| Valence  | 0.3662       | 0.1123       | 0.3399       | 0.1426       |
| Arousal  | 0.3373       | 0.1459       | 0.3581       | 0.1210       |
| Dominance| 0.5291       | 0.0164       | 0.5853       | 0.0067       |
| Dialogues|              |              |              |              |
| Valence  | 0.7520       | 0.1426       | 0.6902       | 0.1971       |
| Arousal  | 0.5474       | 0.3396       | 0.6435       | 0.2414       |
| Dominance| 0.7276       | 0.1635       | 0.7632       | 0.1333       |
|          |              |              |              |              |
|          | Spearman     | p-value      | Spearman     | p-value      |
|          | correlation  |              | correlation  |              |
| Segments |              |              |              |              |
| Valence  | 0.3786       | 0.0997       | 0.3583       | 0.1209       |
| Arousal  | 0.3336       | 0.1506       | 0.3336       | 0.1506       |
| Dominance| 0.5019       | 0.0241       | 0.5584       | 0.0105       |
| Dialogues|              |              |              |              |
| Valence  | 0.8000       | 0.1041       | 0.8000       | 0.1041       |
| Arousal  | 0.4000       | 0.5046       | 0.7000       | 0.1881       |
| Dominance| 0.9000       | 0.0374       | 0.9000       | 0.0374       |

Figure 13: Pearson correlation and Spearman correlation between the human ratings and E-ANEW and human ratings and E-ANEW-TES. Entries are bold if one of the lexicons performed better than the other. The p-value is a 2-tailed p-value that indicates the probability that two uncorrelated datasets produce the same correlation. Most of the correlations are not significant ($p \leq 0.05$). This is probably due to the small number of data points.