Extracting Usability and User Experience Information from Online User Reviews

Steffen Hedegaard
Department of Computer Science, University of Copenhagen
Njalsgade 128, 2300 Copenhagen S, Denmark
steffenh@diku.dk

Jakob Grue Simonsen
Department of Computer Science, University of Copenhagen
Njalsgade 128, 2300 Copenhagen S, Denmark
simonsen@diku.dk

ABSTRACT
Internet review sites allow consumers to write detailed reviews of products potentially containing information related to user experience (UX) and usability. Using 5198 sentences from 3492 online reviews of software and video games, we investigate the content of online reviews with the aims of (i) charting the distribution of information in reviews among different dimensions of usability and UX, and (ii) extracting an associated vocabulary for each dimension using techniques from natural language processing and machine learning. We (a) find that 13%–49% of sentences in our online reviews pool contain usability or UX information; (b) chart the distribution of four sets of dimensions of usability and UX across reviews from two product categories; (c) extract a catalogue of important word stems for a number of dimensions. Our results suggest that a greater understanding of users’ preoccupation with different dimensions of usability and UX may be inferred from the large volume of self-reported experiences online, and that research focused on identifying pertinent dimensions of usability and UX may benefit further from empirical studies of user-generated experience reports.

Author Keywords
User experience; usability; natural language processing; end user reviews; machine learning.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation (e.g., HCI): User Interfaces–Evaluation/Methodology

General Terms
Experimentation; Human Factors; Measurement.

INTRODUCTION
Investigation of a product for usability or user experience (henceforth UUX for short) problems typically requires expensive experimentation. In contrast, informal, but structured electronic word-of-mouth communication between end users affords a potential, cheap source of information concerning UX. Customer reviews on commercial sites such as amazon.com, or dedicated review sites such as epinions.com, contain reviews that are not just summary assessments or recommendations, but also self-reports of the end users experiences, in their own words, in the wild.

The aim of this paper is to quantify the amount of UUX information and dimensions in online reviews from the specific domains of software and video games. We also implement and test a machine-learning-based classifier that tags sentences in reviews according to whether they contain usability or UX-related information and according to the dimensions of usability or UX they pertain to. The primary aim of the classifier is to automatically extract the pertinent vocabulary of end users associated with the various dimensions of UUX. A secondary aim is to investigate the feasibility of using such a classifier to automatically catalogue UUX information found in databases of thousands of reviews, too large for qualified human analysis. We hope to aid the understanding of which dimensions of product use motivate laymen reviewers, and in the future potentially use this understanding when re-designing a product. The scope of the present work is to provide a tool to UUX researchers; future work will explore the automatic identification and extraction of specific actionable outcomes for practitioners.

In order to process information from many different reviews, our approach focuses on extraction of information from individual sentences, rather than entire texts. This is somewhat at odds with approaches in focusing on obtaining a holistic understanding of interaction [16], but as a review may incorporate both good and bad experiences relating to many different dimensions of UUX, we believe that a sentence-based bottom-up approach will yield more precise information about the “typical” vocabulary associated to specific dimensions of UUX.

Related work
User experience has been studied by soliciting user narratives [15, 25, 32] where information is manually extracted from user-generated texts. The volume of texts studied has been substantial (500 texts in [15]), but still small enough for dedicated researchers to process manually, and the users have been specifically asked to write the texts, unlike the typ-
ical online review. Similarly, studies in asynchronous usability testing and reporting [7, 6] have studied user-generated problem reports that are only later reviewed by experts or researchers.

Outside UUX, classification and information extraction from user-generated text is a vibrant research area, both for full texts, and at sentence or short message level, e.g. tweets [39]. Some pertinent examples: Gamon et al. [13] perform sentence-level sentiment analysis on car reviews using several methods from machine learning, but only classify sentences into positive/negative/other. Kim et al. [24] perform sentence-based classification of pros and cons for mp3 and restaurant reviews in order to extract plausible reasons for the reviewers’ recommendations or non-recommendations, but they do not extract vocabulary or classify according to a diverse number of dimensions. Pang et al. [33] classify movie reviews as positive/negative at document level. The primary differences between our work and studies on sentiment analysis as above are twofold: Firstly, we focus on a substantial number of distinct UUX dimensions, some of which may be objective (e.g. a count of false positives from AV software) with no negative or positive opinion, and others which may be described in neutral terms (e.g. aesthetics) where a reviewer can use neutral terms without showing any sentiment. Secondly, unlike most studies in sentiment analysis, the outcome of the classification is primarily a means to an end, namely charting the UUX content of reviews and the vocabulary used by reviewers to describe UUX-related phenomena.

REVIEWS ON THE INTERNET AND UUX

For the purpose of this paper a review is a piece of text detailing pros and cons of a product and possibly an assessment of it and recommendations for potential buyers, written by a user of the product who has been in possession of said product and used it for some time. It may be written either by a professional reviewer or an ordinary end user. We concentrate on reviews assumed to be written by ordinary end users on dedicated web sites, for example epinions.com or amazon.com.

An example of an online review is shown in Figure 1.

Consider the following sentence from a review of the game Gears of War for the Xbox 360:

“You’ll be a little creeped out while playing this game almost all the time.”

The above sentence clearly contains information that is hedonic in nature: Being scared due to the horror elements in the game, and there is a much less clear-element of the satisfaction usability aspect: the sentence communicates a positive experience by the user.

From a communication perspective, user reviews may be viewed as word of mouth communication: informal communication between private parties concerning evaluation of goods and services [1]; reviews from review sites, online fora, and blogs are clearly examples of such informal communication, and are accordingly called eWoM (electronic word of mouth) in the literature [17]. Anderson [1] found that either very satisfied or very dissatisfied customers were more likely to engage in (non-electronic) word of mouth and the word of mouth satisfaction was best described by a bimodal (U-shaped) model. This was confirmed for online reviews by Hu et al. [20] who found that 53% of products on Amazon had a bimodal score distribution (with peaks at very low scores and very high scores). Due to the bimodal distribution, the average score for these products may be misleading. In addition, the information extracted from online reviews may not be indicative of the experience of the average user, but may rather represent those experiences that add or deduct so much from certain users’ experience that they are motivated to write a review.

For (non-electronic) word of mouth communication, extremely dissatisfied customers also engage more in word of mouth than very satisfied customers, though in a sizeable case of their data the differences were not significant [1]; it seems plausible that the same phenomena occur for online user reviews. There is evidence that potential buyers put more emphasis on reviews with low satisfaction than those with high satisfaction as they have a bigger impact on product sales than that of positive reviews and word of mouth [8, 27].

Usability and UX in reviews

Usability is a way to measure a products ability to help a user solve a given task adequately. It is dependent on the product, task, user and circumstances [21], and has been the object of intense academic scrutiny. UX is a younger, emerging field that studies users’ experience with products and the design of
such product with the purpose of generating certain experiences [15, 2].

In usability, traditional studies focus on short term product use (median 30 minutes duration [19]) and conducted in lab settings; few studies stretch across longer time periods and then only weeks [19]. In contrast, most UX research concerns open use situations (61%) and controlled task (33%) experiments, only 20% of papers contain studies based on user-initiated use [2]. No UX research covers longer time periods of months or years which is the expected life span of most products but instead covers at most only a few weeks [2].

In contrast to traditional studies, reviews describe a users opinion and experiences after more protracted use. And in contrast to user narratives solicited for product improvement or research purposes (cf. [15, 25, 32]) online reviews are in a different genre: Authors must follow certain conventions of the review genre, for instance give recommendations on whether or not to buy it. In addition, and unlike narratives written for UX studies, customer reviews on Internet sites appear to be written because the reviewer is motivated by his or her own use of the product, usually in conjunction with some small reward (tangible if the review site offers “credit” for reviews, intangible in the form of community recognition because of the perceived help afforded by a review, or both).

There are important caveats when assessing the potential usefulness of online reviews: It is not clear whether online reviews are written by users typical of the user base; in addition, very few details about reviewers (e.g. gender, age, preferences) are available, in contrast to standard usability studies. Furthermore, some reviews may be fake. Finally, the bimodal distribution of satisfaction present in word of mouth communication leads us to conjecture that in terms of satisfaction, the average user is underrepresented among reviewers, and that reviews may not always yield a representative description of the typical experiences among the user base. However, satisfaction extremes are well represented, it should thus be possible to extract information about situations where the product under review performs both bad and good.

**DIMENSIONS OF USABILITY AND USER EXPERIENCE**

Usability and User Experience are central terms in human-computer interaction. Their precise definition, and their subdivision into dimensions such as Efficiency, Learnability, Hedonic quality, and so forth is still debated [19], in the case of UX hotly so [16, 23, 4], and there seems to be no universal consensus about whether UX is an aspect of usability or vice versa.

We are particularly interested in the way researchers have subdivided UX into various dimensions that pertain to specific aspects, viewpoints, or phenomena within UX. We briefly review existing research below.

**Dimensions of usability**

The ISO 9241 standard [21] defines usability in terms of the three dimensions effectiveness, efficiency and satisfaction:

Usability: The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

A more fine-grained description of usability is obtained by the following five dimensions which, with some variation in the naming of the aspects, are often used for measuring and describing usability in models and literature [12, 36]: (i) **Effectiveness/Errors** [9, 38, 31, 34, 37], 21], (ii) **Efficiency** [9, 21, 38, 31, 34, 37], (iii) **Satisfaction** [9, 21, 38, 31, 34, 37], (iv) **Learnability** [9, 38, 31, 34, 37], (v) **Memorability** [9, 38, 31, 37].

The definitions of the above five dimensions vary somewhat in the literature, and some studies use only a subset of the above [19]. In addition, some studies use a precise and limited definition of measures, and others such as the ISO definition [21] take a broad view of the measures.

**Dimensions of user experience**

Unlike usability, there seems to be much less consensus on the definition of the notion of user experience and its segmentation into meaningful aspects.

The ISO 9241-210[22] definition states:

**User experience:** A person’s perceptions and responses that result from the use and/or anticipated use of a product, system or service.

We may interpret the above as being covered by the **Satisfaction** dimension of the ISO 9241-11 definition of usability [4], but the literature contains many more nuanced interpretations: For example, Bevan [3] describes four dimensions, called **satisfaction measures**, and Ketola and Roto [23] describe a study at Nokia among relevant senior staff who were asked which UX data they found useful, which Bevan later grouped into dimensions [4]. McNamara et al. [29] split the need for evaluation into the three (overlapping) components of functionality, usability and experience. Similarly, Hassenzahl [16] divides UX analysis into the three partially overlapping approaches beyond the instrumental, emotion and affect, and the experiential, but in later work [14] shifts focus to the subjective side of product use and relates it to Self Determination Theory [35] and Flow [11]. Bargas-Avila and Hornbæk [2] systematically collect a sample of 51 publications from 2005-2009 reporting empirical studies on UX and describe a number of (non-mutually exclusive) dimensions together with the percentage of studies from their sample that pertain to the identified dimensions.

**Dimensions selected for this paper**

Based on the literature above, we elected to use the 5 standard dimensions of usability, and chose the sets of dimensions from 3 of the studies of UX that had both precise definitions of the dimensions and clear demarcations of the differences between them (with two exceptions in FREQUENT, see below).

A summary is shown in Table 1; we briefly describe the dimensions below.
Dimensions of CLASSICUA [12, 36] are: 

**Errors/effectiveness:** The number of (non-fatal) errors made by users on their way to completing a task or the quality of task outcome. 

**Efficiency:** The speed or other measure of cost associated with performing the task for users at a given experience level. 

**Satisfaction:** A subjective rating of satisfaction with product use or liking of the product or features. 

**Learnability:** The amount of time it takes to learn to use the system, how difficult it is for a first time user or development over time. 

**Memorability:** How well users retain information gained about or through the system. 

The dimensions of BEVAN [3] are: 

**Likability:** The extent to which the users are satisfied with their perceived achievement of pragmatic goals, including acceptable perceived results of use and consequences of use (note the close similarity with Satisfaction from CLASSICUA). 

**Pleasure:** The extent to which the users are satisfied with their perceived achievement of hedonic goals of stimulation, identification, evocation and associated emotional responses. 

**Comfort:** The extent to which the users are satisfied with product use or upgrading from a previous product. 

The dimensions of KETOLA [23, 4] are: 

**Anticipation:** What did users expect, what is the anticipated use? 

**Overall usability:** Was the user successful in taking the product into use or upgrading from a previous product. 

**Hedonic:** Fulfillment of inner needs such as pleasure, enjoyment, or things preventing this such as frustration. 

**Detailed usability:** Going into details on which functions are used, ordinary usability problems and performance satisfaction/problems. 

**User differences:** Differences between users such as previous product experience, how they access features and differences between the actual buyers and target user group. 

**Support:** Aspects with the human- or software-support available and how it affects user satisfaction, possible product returns, or user wish lists. 

**Impact of use:** If and how the new device change the usage patterns of the users. 

The dimensions of FREQUENT are: 

**Affect and Emotion:** Affect and emotion induced by using the product, including other aspects such as Enjoyment, fun and Frustration. This dimension fully encompasses Enjoyment, fun and Frustration, and would be considered encompassed by the Hedonic dimension. 

**Enjoyment, Fun:** How entertained is the user while using the product? This is also an affect and emotion, and hedonic, dimension. 

**Aesthetics, Appeal:** Appreciation of beauty or good taste. Typically associated with graphics or sound. 

**Engagement, Flow:** How engaged is the user in using the product forgetting everything else? Also includes challenge versus skill balancing needed for achieving flow state. 

**Motivation:** What motivates the user in using the product (task/inner motivation etc.)? 

**Frustration:** Frustration or hardship induced by using the product. This is also a negative hedonic dimension. 

We discarded the two dimensions Generic UX and Other as reported by [2] as no clear definition was available.

**PRE-STUDY**

To see whether randomly sampled reviews contain sufficient UUX-related information to warrant further study, we performed a pre-study among usability experts who were given a sample of Internet reviews and a free-form exercise asking them to mark sentences containing information about usability or UX.

**Participants**

9 usability experts were contacted, all of whom are active researchers in usability (7 from academia, 2 from industry). Of these, 8 gave affirmative answers and were enrolled as participants. All participants were compensated with two bottles of wine.

**Procedure**

24 reviews were sampled on January 5 2012 from 12:00 – 14:00 from the website epinions.com, collecting the 6 most recent reviews of each of the four categories “Digital Cameras”, “Headphones”, “Software” and “Video games”. 3 reviews were discarded, all from the software category (1 was a review of an iPhone game (games were covered explicitly in another category), and 2 were reviews of printed children’s books). Each review was randomly assigned to 2 distinct participants.

Each participant was asked to read and comment on six different reviews in total. The participants received only written instructions asking them to free-form annotate text in the reviews that they found interesting concerning (a) their own perception of usability, and (b) user experience. Participants were neither given definitions of usability or user experience, but encouraged to use their own perception of these terms. Each participant was asked to use at most 2 hours in total on all six reviews, including time to read and to annotate.

We collected the annotated texts and post-processed them in two ways:

**Raw containment of UUX:** Each review was manually split into sentences and was marked with the identity of a participant if the participant had marked part of or the entire sentence as relevant. Due to the level of annotation performed by most experts, no distinction was done between dimensions or UX and usability based on the experts comments.
Presence of UUX dimension: For each review, the first author coded all sentences annotated by at least one participant using the dimensions from Table 1; the same sentence could be annotated with more than one dimension. Examples of dimension assignments can be seen in Table 2.

<table>
<thead>
<tr>
<th>Content</th>
<th>Dimensions present</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you like multiplayer strategy games, buy this with confidence.</td>
<td>satisfaction, user differences anticipation, satisfaction</td>
</tr>
<tr>
<td>Those expectations were met. Mostly, anyway.</td>
<td>user differences, flow, enjoyment, hedonic satisfaction</td>
</tr>
<tr>
<td>... making the game enjoyable for beginners as well as veterans.</td>
<td>satisfaction, enjoyment, hedonic engagement/flow</td>
</tr>
<tr>
<td>Multiplayer is excellent, but the single player campaign isn’t.</td>
<td>satisfaction, frustration</td>
</tr>
<tr>
<td>Most of the missions are enjoyable, and each one has optional goals which add replay value.</td>
<td>satisfaction, enjoyment, hedonic engagement/flow</td>
</tr>
</tbody>
</table>

Table 2. Examples of annotation from the video game Starcraft 2.

Results

Raw containment of UUX: Calculation of inter-rater agreement for raw containment of usability or UX indicated that participants were somewhat in agreement on which sentences did, or did not, contain any relevant information at all, but that no hard conclusions should be drawn based on the data (Krippendorff’s $\alpha = .783$)\(^1\).

In total, 13% of all sentences were marked as relevant to usability or UX by both participants assigned to each review, 36% as relevant by one, but not both assigned to each review, and 51% of all sentences were unmarked (i.e., deemed as irrelevant by participants).

There was great diversity in the understanding of UUX and annotation volume per participant. One participant specifically noted that he had given up marking user experience data as it “virtually encompassed everything”, and only a single participant consistently annotated UX and usability information as two distinct categories. We observed some discrepancies in annotations; for example, one participant had marked the sentence “The product works extremely well” as relevant for UUX in a review, but later in the review failed to mark the similar sentence “It also works well when listening to music while using power tools . . .” as relevant.

Presence of usability or UX dimension: The results are summarized in Table 3.

For the classic usability measure as seen in Table 3, almost all sentences in the dimension errors/effectiveness were describing quality of task outcome (e.g., music quality for headphones), but a few classic error counts were also present (e.g., notes correctly transcribed by a sheet music scanning feature of a program, and false positives in anti virus software).

Only rarely ($N = 4$) did reviews attach any numbers to measures of efficiency and effectiveness, and even then they were not considered as exact measures, but merely rough estimates such as “...and the whole process only takes about 5 minutes.”

Detailed inspection of the reviews revealed that some dimensions only occurred in specific product categories: The dimension physical comfort was exclusively encountered in camera and headphones reviews, and the dimension pleasure mainly for video games.

The dimensions of motivation and enchantment, both popular dimensions in empirical user experience research being represented in 8% and 6% of papers respectively [2], were not encountered at all in the pre-study.

![Table 3. Occurrences of dimensions found in sentences annotated by participants as a percentage of the total number of sentences in all reviews.](image)

In summary, 13% + 36% = 49% of all sentences were marked as relevant by at least one of the two participants annotating each review. Some confirmation bias may be present as participants were specifically asked to look for information relevant to usability or user experience, but based on the results we concluded that the volume of text in a review relevant to usability or UX, and the dispersion of text across UX dimensions were both substantial enough to warrant a larger-scale annotation experiment.

FIRST STUDY: ANNOTATION OF REVIEWS

Based on the promising results of the pre-study, we opted to harvest a larger sample of reviews and annotate them. We decided to keep the per-sentence annotation of the pre-study and concentrate on only 2 product categories as it would allow us to annotate more sentences in each product category while retaining the ability to make comparisons across categories.

Procedure

We collected reviews from the two product categories Software (520 reviews) and Video games (2972 reviews across various PC and console platforms) on the epinions.com website on July 5th, 2012. All public available reviews in the two categories were collected. We split each review into sentences using a routine from the Python NLTK [5] which came pre-trained on the British National Corpus. We then drew sentences randomly from the pool of all sentences above and performed manual annotation on each drawn sentence. In total 4587 sentences (of a pool of 132609) were annotated from the Video games and 611 (of a pool of 18646) from the Software.

![Table 4. Word count statistics for reviews.](image)

The annotation was conducted by one of the authors and two graduate student annotators: The graduate students were given a short, written introduction description of all dimensions and participated in a co-annotation workshop for four
hours with one of the authors acting as instructor and senior annotator. This was followed by four hours of individual annotation where the student annotators could freely consult the senior annotator for questions. Inter-rater agreement was computed at the end of the individual annotation with Krippendorff’s $\alpha$ for all dimensions in the $[0.9 - 1.0]$ range. Each graduate student annotator then continued individually for 22-25 hours over the course of two weeks, and the senior annotator for 10 hours during one week.

All annotations were performed in a custom-built tool built by a research programmer not otherwise involved in the study. A total of 6655 sentences were annotated with 1315 sentences annotated by the senior annotator 2922 sentences by annotator 2, and 2418 sentences annotated by annotator 1, every tenth sentence annotated by an annotator was randomly chosen from sentences already annotated by another in order to compute inter-rater agreement; all dimensions annotated showed a high degree of agreement (Krippendorff’s $\alpha$ lowest score $\alpha = .842$; 22 of 25 dimensions had $\alpha > .90$).

Some examples of sentences from reviews and their annotations:

- “Once again, the way sound distorts during the slow-motion sequences adds a nice touch to the experience.”. This sentence mainly describes Aesthetics and subsequently the effect on Engagement while playing the game, yet also expresses Satisfaction with the effect. As this example illustrates, several dimensions are often encountered together with Satisfaction.

- “The sound is what you’d expect from a Nintendo title, and can become quite annoying after extended plays.” This sentence describes displeasure, relevant to the dimension Pleasure; this is also a measure of Satisfaction with the sound.

Results

Table 5 shows the fraction of sentences in each product category annotated by the various UUX dimensions.

The table confirms the observation from the pre-study that some dimensions are hardly used at all: In CLASSICUA, Memorability does not occur at all in the software category, and only in 0.37% of sentences among video games. Likewise, Efficiency occurs very rarely (4.45% in the software category, 1.12% in video games). Among the dimensions from BEVAN, Comfort and Trust are again absent, but Likeability is prominent, as expected from the pre-study. The dimension Pleasure is more prevalent among video game reviews than reviews of software.

In KETOLA, Impact is almost completely absent, and Support rare, but more prevalent among software products, possibly reflecting that this dimension is more valued among users of software products than video games. Conversely, the dimension Hedonic is present in 7.77% of all sentences sampled in the video games category, more than twice as often as in the software category.

It is striking that there are quite few sentences annotated by dimensions in the FREQUENT classification (9.25% for software, 29.72% for video games) compared to the CLASSICUA, BEVAN and KETOLA categories (where more than 40% of all sentences contain usability or UX information). This phenomenon is due to the FREQUENT classification’s lack of a “catch-all” category such as CLASSICUA’s Satisfaction, BEVAN’s Likability and KETOLA’s Detailed usability.

In summary, the results support two clear conclusions: First, the product domain of the review (in our case, software, resp. video games) influences the amount of sentences that pertain to specific dimensions of UUX (e.g., the dimension engagement, flow is much more prevalent in reviews of video games than in reviews of other software). Second, the four sets of dimensions of UUX we consider differ very much in the balance of dimensions: Clearly, FREQUENT, the model containing the most categories, has the most even distribution of sentences across the various dimensions, whereas the other sets of dimensions seem to have greatly skewed distributions towards the “catch-all” categories described above.

SECOND STUDY: AUTOMATIC CLASSIFICATION OF SENTENCES IN UUX DIMENSIONS

Based on the results of the first study, we wish to investigate the vocabulary employed by users when conveying information relevant to the dimensions of UUX. One straightforward way of extracting such a vocabulary is to construct a machine learning classifier that discriminates between dimensions based on words or other features of the text that are automatically computed during the training of the classifier. An added benefit is that the constructed classifier, if precise, may be used for automatically tagging a sentence with the UUX dimensions it pertains to. This tagging task may be viewed as a set of binary classification tasks: For each dimension, and for each sentence, does the sentence pertain to that dimension, or not? Such a tagger may be used to either aid future researchers in manual annotation, or in lessening the amount of sentences to be studied (i.e., sentences receiving no tags by an automatic classifier can be ignored at little risk).

Procedure

For each UUX dimension a binary classifier using a bag-of-words feature set was trained and evaluated in a sequence of steps as follows: Preprocessing step: Each sentence was tokenized, words from the NLTK stop word list [5] removed and the remaining words stemmed using the Snowball stemmer. Data split step: The total dataset is split into a five-fold stratified cross-validation [28] scheme. Training step: For each cross-validation split, the training set is used for creating feature vectors with TFxIDF weighting and $\chi^2$ [28] ranking is subsequently used to discard the 10% worst discriminating features. A linear kernel Support Vector Machine (SVM) [10] is then trained using the feature set. Classification and validation step: For each cross-validation split, classification of the evaluation set using each SVM classifier is performed, and the classification results for all the cross-validation splits are aggregated, and standard performance measures (see Table 3)
Recall 2

Table 5. Distribution of dimensions in sentences within Software and Video Games reviews. The “Any dimension” rows indicate the percentage of sentences annotated with at least one dimension. Each sentence can be annotated with more than one dimension, hence “Any dimension” is not the sum of the other numbers in the same column. Differences between the Software and Video Games categories were tested for significance using the non-parametric two-tailed Wilcoxon rank-sum test [18] and significance at p < .05 is indicated in boldface.

Table 6. Definitions of precision, recall and F1.

Table 7. Significant results are marked in bold.

Table 8 holds the 30 most important word stems for each dimension where the difference in precision, recall and F1 between the classifier and the baseline was significant. As an example of top word stems associated with a dimension not included in the table are “sooth”, “cute”, “reliev”, “handhold” and “exist”, all associated with the dimension trust.

For the dimension Frustation with word stems “frustrat”, “incompatibilit”, “hardest”, “perpetu”, “insult” what seems like less relevant words also made it to the top 30, for instance “babysit”. This is due to reviews containing text such as “And that is of course an AI partner controlled friend, there is nothing that can ruin a good RPG then a partner that is supposed to be helping you but instead makes you feel like your babysitting a 5 year old with mental problems, on top of battling blood thirsty monsters.” This particular sentence also illustrates the intricacies of our task: Clearly, the use of “babysit” is figurative, not literal, hence signals frustration.
Other spurious word stems in Table 8 (e.g., “sc”, “thus”, “gps”) can be attributed to two phenomena: (i) the word stems in the dimension have low discriminatory power, whence the classifier was barely able to distinguish relevant/irrelevant sentences, and (ii) a word stem may be considered important if it by chance occurs in the training corpus of the classifier in a small number of sentences, all of which are relevant to the dimension.

Dimensions such as Hedonic, Pleasure and Affect and emotion which fully or partially encompass other dimensions tend to share most of the top places of the included subcategories. For example, nine of the ten most important word stems from the dimension Enjoyment and Fun were found among the top 15 important word stems in the encompassing dimension Effect and Emotion which also rated other word stems such as “scari” and “frustrate” related to emotion, but not enjoyment highly. The notable exception to this phenomenon is the dimension Detailed usability that encompasses all measures of classic usability as well as mention of specific usability problems; this dimension has many highly ranked word stems relating to Satisfaction, but fewer highly ranked words stems also occurring among the other classic usability dimensions.

### DISCUSSION

We have unearthed a significant difference between software and video game reviews in terms of which UUX dimension they frequently mention. With the exception of satisfaction software reviews emphasize classic usability measures more than video game reviews, which in turn put more emphasis on dimensions such as Hedonic, Affect and emotion, Pleasure and Enjoyment and fun. The notable exception to this difference is Frustration for which the difference between software and video games were not significant.

The automatic classifier works well on commonly occurring dimensions, but tends to be too conservative even for these dimensions. There are ways of improving such classifiers, but our experiments suggest that simple off-the-shelf machine learning solutions are insufficient. When sifting through large amounts of material, it is easy to miss infrequent, but potentially important information, for instance the presence of information pertaining to Enchantment or Trust, and the classifier clearly is unable to assist a human expert in this regard. However, for some of the commonly occurring categories, the classifier may assist by removing sentences irrelevant for the dimensions, at the cost of some potentially relevant sentences being removed.

The set of word stems extracted shows that some dimensions (e.g., Enjoyment, Fun) have associated vocabularies containing words closely related to the words used to describe the dimensions in the literature (e.g., word stems that are synonymous or antonymous to enjoyment and fun, describing respectively positive and negative experiences), but other dimensions such as Errors/effectiveness instead have vocabularies related to specific problems and errors such as “lag”, “glitch”, “imprecise” and “bug”. This illustrates the varied vocabulary used by reviewers when describing specific dimensions of interaction, and that the vocabulary is more varied for some dimensions than others.

Our results strongly suggest that more complex information about how users express their feelings and experiences with situations and problems related to UX can be extracted from reviews and other narratives. The results also suggest that the task of mapping the users’ utterances to specific dimensions of UX is only partially possible to do in an automatic fashion and that some of these dimensions are associated to a complex vocabulary.

When using sets of dimensions of UX for practical purposes—for example for gauging which dimensions of a product are perceived to be important by users—then sets with many complementary dimensions such as FREQUENT appear to be more fine-grained and informative than other such sets. It may be possible for the UX community to settle on a small, possibly domain-dependent set of dimensions, simply by performing empirical investigations such as ours, or in more traditional settings such as usability tests.

### LIMITATIONS

While our results shed light on the general UX concerns of end users, the anatomy of the reviews we considered, possibly Internet reviews in general, does not seem to contain much detailed information about specific situations of use, or of measurements. No reviewer writes “The number of mouse clicks to navigate from the start screen to the functionality I want is 7, and that this is annoying”. While it is conceivable that a professional software reviewer, or an end user taking a conscious interest in usability, would write such a sentence, we did not encounter these. Thus, it seems highly unlikely that mining Internet reviews can supplant traditional usability testing or UX studies.
Finally, the sentence-based annotation has acted as a convenient proxy: If a user spends 10% of the sentences in a review discussing matters related to Enchantment, it is likely evident that enchantment is a major part of his view of the product; but there may be other measures that more precisely reflect how much the user is occupied with different dimensions of UX.

Future work
The data we considered were limited to two specific domains (software and video games), and the volume of data, while respectable, was insufficient to establish a vocabulary for all usability dimensions. Future studies must extend our work to more domains, and must consider a very large volume of data (a rough estimate based on our work: several tens of thousands of sentences). In addition, the idea of extracting vocabularies and associating features of texts or other utterances to dimensions of UUX can be applied to other domains, including spoken words at traditional lab-based usability studies. It seems worth to investigate the difference across product domains of the distribution of sentences among UUX dimensions found in this study. Likewise, it is interesting to link the presence of sentences pertaining to the UUX dimensions to attributes of the reviews that can be inferred otherwise, for example negative vs. positive reviews, or the helpfulness of reviews as voted upon by other users.

Filtering and grouping of dimensions may be examined in greater detail in follow-up studies also investigating actionable outcomes. Finally, a better-performing classifier, or human annotation of a larger amount of complete reviews instead of isolated sentences may allow for analyzing the distribution of the dimensions we consider, on a per-review basis.

REFERENCES


