Spring 5-29-2015

Who Are We Listening to? Detecting User-generated Content (UGC) on the Web

Marc Egger
University of Cologne, egger@wim.uni-koeln.de

André Lang
Insius UG, andre.lang@insius.com

Detlef Schoder
University of Cologne, schoder@wim.uni-koeln.de

ISBN 978-3-00-050284-2

WHO ARE WE LISTENING TO? DETECTING USER-GENERATED CONTENT (UGC) ON THE WEB

Complete Research

Egger, Marc, University of Cologne, Cologne, Germany, egger@wim.uni-koeln.de

Lang, André, Insius UG, Cologne, Germany, andre.lang@insius.com

Schoder, Detlef, University of Cologne, Cologne, Germany, schoder@wim.uni-koeln.de

Abstract

The analysis of text-based user-generated content (UGC) on the Web has become one highly acclaimed topic in recent years both in theory and practice. As users are able to participate and publicly comment on almost any webpage nowadays, UGC occurs scattered across the web and mixes with various content types such as advertising texts, product descriptions or other editorial articles. Holistic research that aims to listen to the voice of the consumer therefore needs to separate UGC from non-UGC. Unfortunately the UGC characteristic is not a directly observable attribute of content. As the amount of public available textual data on the web is vast and increases rapidly, manual classification is not applicable in this "big data" environment. From this, the previously unmet need emerges to perform UGC classification automatically, for which we provide three contributions. First, we show that UGC incorporates signals that enable humans to context-free decide whether a text has been written by another user. Second, we show that these signals can be utilized by supervised machine learning to perform UGC classification automatically. Third, we demonstrate and evaluate the fundamental feasibility of our approach on a dataset of German language web texts.

Keywords: user-generated content, machine learning, text classification, support vector machines, SVM

1 Introduction

Web 2.0 and social media have enabled consumers to express themselves online. Consumers share and discuss their thoughts, opinions, experiences, and feelings online while creating a vast, publicly available data source (Egger and Lang 2013). Online comments published by users are typically termed user-generated content (UGC) and can include various content types. Beside the rising quantity of audio-, video- and image-based content, a huge amount of UGC on the Web is still text, which is the subject of investigation in this article. Within UGC, consumers discuss general topics as well as brands and products. Because consumers also search for online discussions and utilize online consumer reviews for consumption decisions (Park et al. 2007), UGC not only affects non-customers' perceptions but also has the potential to influence future customers (Decker and Trusov 2010), which has a direct impact on commercial interests. Given this commercial potential, it is no surprise that organizations are interested in silently listening to consumers while deriving insights from this public data source for market research, marketing activities, and product or brand management. Likewise, academics have also realized the potential of distilling information from UGC. The number of contributions in the marketing literature and opinion mining—an interdisciplinary research field incorporating text mining, information retrieval (IR), and sentiment analysis methods—has increased significantly over the past years.

Despite the enormous potential of UGC analysis, the challenge of automated UGC detection outside the Web remains open as participation within the global Internet community is basically unrestricted

and democratized and has low entry barriers. Private persons as well as institutions, organizations, and enterprises can easily contribute and wire their created content to the global repository of connected data. Individuals can either establish their own channels (e.g., websites, blogs, forums, or even their own platforms or social media communities), or participation can happen on preexisting channels (e.g. websites that enable participation via comments, product-review platforms, forums, or social media communities). Thus, UGC on the Web is as decentralized as the Web itself and often mixes with non-UGC (e.g., editorial content). Unfortunately, content does not incorporate a directly observable attribute that classifies it as UGC.

As both theory and practice often target the analysis of UGC (e.g., in voice-of-the-customer analysis, reputation management, or brand/image analysis), there is an important gap in terms of how to find and separate UGC from other content found on the Web. The typical and less-than-ideal solution of restricting research to certain data sources for which the presence of UGC is assumed (e.g., a special website, forum, or product-/movie-review community) leads to several problems. First, a pre-selection bias cannot be excluded when restricting research to specific data sources, such as product-review platforms. Furthermore, it can also be assumed that the above-stated approach neglects a substantial amount of the total available UGC.

With this article, we aim to make three contributions. First we show that regardless of context, consumers are able to decide whether a Web text is UGC or editorial content. Second, we propose a supervised machine-learning method that aims to semi-automatically detect UGC. Third, we provide a study whereby we show the fundamental feasibility of automatic UGC detection and evaluate our approach within the domain of online synchronization services. Therefore, we attend to the previously unaddressed need to automatically separate UGC from non-UGC to enhance analysis approaches directly targeting the measurement of the "online consumer voice". This has great and increasingly practical relevance for organizations' brand, reputation, and public relation management and can further expand the scope of previous research focusing on a subset of UGC (e.g., forum posts or product reviews).

The remainder of this article is structured as follows. After providing a definition for and outlining our understanding of UGC, we investigate whether consumers are able to decide whether a text is UGC or editorial. After this, we discuss the characteristics of UGC and editorial texts before we describe our supervised machine-learning approach. Finally, we evaluate this approach on a self-created gold set of German language Web texts, discuss the limitations to our research, and provide an outlook to future research.

2 User-Generated Content

Before we turn to UGC detection, we provide definitions for and outline our understanding of UGC, which is often referred to within the scope of Web 2.0 and social media. Even though the term UGC is neither fixed nor selective (Grob and Vossen 2007), one of the most quoted definitions of UGC is provided by the Organization for Economic Cooperation and Development (OECD) (Vickery and Wunsch-Vincent 2007). OECD uses the term user-created content (UCC), which is considered synonymous with UGC. According to Vickery and Wunsch-Vincent (2007), UGC has three central characteristics: (1) publication requirement, (2) creative effort, and (3) creation outside professional routines and practices. The publication requirement notes that content has to be generally accessible to an unrestricted group (e.g., on a webpage) but, at a minimum, to a certain group of people (e.g., within a social network). This constraint thus excludes all content and communication that are used for directional (i.e., one-to-one) communication (e.g., e-mail, instant messenger, or similar services). The second requirement—creative effort—relates to the necessity of a minimum amount of personal contribution to create the content. Therefore, copied content does not count as UGC, while collaboratively created content, (e.g., Wikipedia articles) does. The third characteristic requires content to be created outside professional routines and practices. This means that content creation must not be motivated by monetary incentive. Therefore, this requirement separates UGC from content that has been created within a

professional and/or commercial context, which is often referred to as editorial content. This characteristic cannot always be selective as platforms sometimes financially incentivize user content creation (e.g., Ciao.com). These rewards are acceptable as long as the primary motivation to create content is still reputation, recognition, contact with other people, or self-fulfillment (Vickery and Wunsch-Vincent 2007).

Most other descriptions and definitions of UGC share the characteristic of content being created outside professional routines and practices and provide additional specifications, restrictions, or limitations. For example, Bauer (2010) extends the publication requirement to include self-publication and the absence of editorial pre-selection. Thus, the author limits UGC to include only electronic media content that has been deliberately created and made accessible to the public on the Internet by Internet users and that has not been editorially pre-selected, created within a professional environment, or published for the sake of commercial purposes (Bauer 2010). This definition might be criticized because it is more restrictive and because some Web channels are required to perform at least minimal pre-selection or erase legally dubious content (Feldkamp 2007). As an example, some newspaper websites that provide the ability to comment on articles are required to check user comments before publication (Hermida and Thurman 2008). Because many other contributions to the literature use definitions for UGC that share characteristics of the OECD definition (Daugherty et al. 2008; Grob and Vossen 2007; Kaplan and Haenlein 2010; Krumm et al. 2008; Weimer et al. 2007), we also refer to the OECD definition (Vickery and Wunsch-Vincent 2007) as a basis for our article.

Finally and referring to the above-stated discourse, it is important to note that UGC relates to a certain role for the communicator as he/she publishes content. This role is usually taken on within a private context without professional or commercial incentives and with the motivation of creating content for reputation, contact with other people, or self-fulfillment.

3 Related Research

To our knowledge, prior research has surprisingly not addressed the detection and separation of UGC versus non-UGC—a topic located within the research fields of information retrieval, natural language processing (NLP), and text mining.

Because consumers often express their personal points of view to a certain subject through UGC, related research can be expected on the topic of opinion and subjectivity detection, which is usually investigated within the research fields of opinion mining and sentiment analysis (An overview can be found in (Pang and Lee 2008). Both subjectivity and opinion detection are often targeted by modelling the problem as a binary text-classification task, which is addressed using machine-learning approaches. The detection of opinion texts was the focus of the 2006 blog track at the Text REtrieval Conference (TREC) (Macdonald et al. 2010). Here, Zhang et al. (2007) propose a system to detect opinions within blogs by utilizing support vector machines (SVMs) as a classifier. Similarly, Huang and Croft (2009) propose a system that automatically adds additional keywords (i.e., keyword expansion) to a user-defined keyword Web search that hints to subjectivity within a certain topic.

It is important to note that UGC, opinion, and subjectivity are related, but while the latter two *describe the semantic attributes of a comment,* UGC *targets the role of the communicator*. This means that opinion and subjectivity can be communicated from a user within a private context (i.e., UGC) as well as from a professional or commercial perspective (i.e., non-UGC). Although UGC can be expected to carry a great deal of opinion and subjectivity, the relationship is not conclusive. Therefore, the segmentation criteria of subjective opinion texts differ from our research purpose in that, here, the semantic meaning of texts is determined and not the role of the communicator. However, the applicability of semi-supervised text classification shows that cognate subjects can be addressed by these kinds of methods. The same applies to the fields of text-genre detection and classification, for which different focal genres (e.g., fictional versus non-fictional (Kessler et al. 1997) or informative versus imaginative (Karlgren and Cutting 1994) are addressable by text-classification approaches.

While the above-stated contributions all refer to the semantic level of texts but not the communicator him-/herself (e.g., author), researchers have also aimed to detect attributes not contained in the written text itself but instead manifesting in detectable characteristics of the text, such as the type of author or the author him-/herself, which matches to the purpose of the article at hand. Text author's gender detection (Burger et al. 2011; Corney et al. 2002; Koppel et al. 2002; Yan and Yan 2006) or sock puppet detection (Solorio et al. 2013) are examples of contributions that aim to discover the role and/or attributes of the author. The detection of the author him-/herself is investigated within the research field of authorship attribution (Stamatatos 2009). For example, Diederich et al. (2003) utilize SVMs to identify authors of several German newspaper articles with a high reliability of 60-80%. For authorship attribution, single authors are the subject of investigation. However, for our research purposes, we aim to detect a certain group of authors, all of whom share the attribute of being in the role of "user" according to the definition of (Vickery and Wunsch-Vincent 2007).

Because the above-stated problems are usually mapped to text-classification tasks addressed by machine-learning methods, such as Naïve Bayes, SVMs, or conditional random fields, these kinds of tools are likely to be promising approaches for UGC classification.

4 Decidability and Attributes of UGC

Because the distinction between UGC and non-UGC is drawn based on the process of its creation rather than the characteristics of the produced content itself, it is usually not directly observable. Although sources like product-review websites suggest that the content is supposed to be written by users outside professional spheres and should therefore be classified as UGC, this is merely a best guess based on the type of website, text characteristics, or other latent indicators. Still, the author and his/her motivation to create content remains unknown. This is especially for true for blogs, for which it is almost impossible to draw a clear line between professional and private creation routines by merely looking at the results. Despite the widespread use of the term UGC and the academic usage of content denoted as UGC, little distinction has been made about this subject so far.

While the true attributes of UGC are mostly unobservable, one may presume that the different kinds of authors, motivations, and routines are reflected in the resulting content. Therefore, finding methods to comprehensibly infer whether text is UGC or non-UGC from latent information (e.g., text characteristics or place of publication) is an indispensable task.

The differentiation between UGC and non-UGC can be made based on any information present and available at the time the content is retrieved from the Web. This information can be divided into two classes: attributes inherent in the textual content itself and contextual information. Textually inherent attributes relate to the style of writing, the use of words and language, grammar, and stylistic and syntactic attributes (e.g., punctuation). Contextual information is any other information besides the text itself and can include the website's name, URL, page design, meta-information, and any other possible indicators present on the webpage (e.g., a "posted at:" header or the author's pseudonym).

Since the number of possible contextual indicators is vast and meta-information (e.g., posting date, pseudonyms) of online articles is very heterogeneous, the proper selection of relevant features for UGC classification remains difficult. On the other hand, textually inherent attributes are the lowest common denominator, so any differentiation could be uniformly applied as long as text representation is available. Therefore, using textual inherent attributes as the basis for UGC classification is especially promising for an approach that aims to maximum applicability.

To investigate whether those textual inherent attributes exist, we first have to verify if human test subjects are able to consistently determine whether, regardless of context, a piece of Web text has been published by a user in line with the above-stated definition.

Therefore we collected textual content from the Web using public search engines as these usually serve as the entry point for users to discover content of interest, which can be both UGC and non-UGC. As a topic, we chose the synchronization service MobileMe from Apple, which is the predeces-

sor of iCloud and has been discussed very thoroughly by consumers in past years. To gather data, we developed a Web crawler that collects webpages using general purpose search engines, such as Google, using the keywords "mobileme" or "mobile me." We collected 5,520 webpages from the search engine Google and extracted 218,438 continuous text blocks from which we removed boilerplate elements, such as side navigation or link lists using the approach from Kohlschütter et al. (2010).

To study decidability, we randomly selected 500 text blocks, which we denote as "documents" in the following. To determine the feasibility of manual classification, we instructed three test subjects on the characteristics of UGC according to the OECD definition (Vickery and Wunsch-Vincent 2007) and asked them to independently classify the 500 documents into UGC versus non-UGC. Two of the test subjects can be considered experts at handling UGC, while the third test subject is an average Internet user without specific background knowledge.

To determine the effectiveness of the approach, we first have to test whether the characteristics of UGC can be objectively concluded from the dataset. If test subjects evaluate this characteristic very differently, an Information Retrieval (IR) system will not be able to independently and universally produce valid results, in which case, this basic problem of IR is not decidable (Manning et al. 2008). To prevent this issue, we measure the degree of agreement across different test subjects, which is known as inter-rater reliability. Inter-rater reliability for a three-person test is usually tested using Fleiss' multirater kappa (Randolph et al. 2005). However, the application of Fleiss' multirater kappa requires that test subjects have guidelines regarding how many elements have to be assigned to each class. Therefore, Randolph et al. (2005) propose free-marginal multirater kappa as a variant, which does not have a precondition for the number of assignments to each class. We use free-marginal multirater kappa for our purposes, for which the kappa metric is defined within [-1;+1], with 1 denoting a perfect match, 0 equaling a stochastic hit, and -1 describing a below-stochastic match.

In our study, we achieved an agreement of 84% and a kappa of 0.680. Notable are the differences between the two experts (Test Subjects A and B) and the ordinary Internet user, whose decisions regarding UGC sometimes differed from those of the experts (Table 1). Nevertheless, the results led us to the conclusion that the classification of UGC based on textual inherent attributes seems to be an objectifiable characteristic.

	A	**B**	**C**	**Sum**
A	-	0.880	0.824	0.852
B	0.880	-	0.816	0.848
C	0.824	0.816	-	0.820

Table 1: Test persons agreement relating to UGC classification

4.1 Textual Attributes as a Basis for UGC Detection

Based on our study of letting test subjects decide whether a text has been written by a consumer or from a professional perspective, we conclude that UGC has textual attributes that differentiate it from non-UGC. The description of those attributes relates to the research field of textual linguistics. Textual linguistics is concerned with linguistic structures and formulations within texts (Vater 2001). One of the tasks within textual linguistics is classifying texts into text types. Humans are able to recognize different types of texts as well as perceive and apply inherent rules (Vater 2001). Examples of text types are official letters, journalistic articles, e-mails, guest books, and newsgroups (Stein 2002).

There are different criteria that enable individuals to assign texts to certain text types (for an overview, see Vater (2001)). Which criteria are chosen depends on the objective of the classification. Sandig (1972) proposes an assignment according to attributes of ordinary language texts for the German language. This classification based on attributes has parallels to the topic of text classification within the context of information retrieval, which we will discuss later. According to Sandig (1972), two out of three of the most important attributes are the differentiation of monologues and dialogues as well as

spontaneous and non-spontaneous texts. In addition, certain text types usually incorporate introduction or finalization formulations (e.g., personal forms of address at the beginning or salutations at the end of a letter or e-mail). Other attributes relate to the choice of words and syntax. Texts can be written rather formally or informally, the latter resembling spoken language to a greater extent and often being characterized by a colloquial wording and a more lax use of orthographic rules and abbreviations (Stein 2002).

Based on feedback from the test subjects of our first study, we propose additional attributes, such as the use of self-referential personal pronouns (e.g., "I," "me") or a second-person form of address (e.g. "you"), which is often found in a social Web context. Furthermore, the use of acronyms like LOL ("laughing out loud") and ROTFL ("rolling on the floor laughing") as well as subjective words like "super," "fine," or "bad" can be found in UGC. In contrast to this, in the German language, test subjects reported the use of more formal addresses (e.g., "Sie") within non-UGC.

Table 2 shows an overview of the text types on the Web according to these different attributes. The last five attributes seem to provide adequate differentiation criteria for UGC. We must also note that there are at minimum two exceptions: Because wikis are rather different from the other UGC text types, the proposed criteria seem to be unable to differentiate wikis from editorial encyclopedic texts. The second exception is advertising texts, which often include subjective, personal, and informal speech.

	Text Type	Monologic	Spontaneus	Introduction form.	Finalization form.	Informal	Pron. 1 (I)	Pron. 2 (You)	Internet slang	Subjective
UGC	Blog	+/–	–	–	–	+/–	+	+	–	+/–
	Comment (Blog, News)	–	+	+/–	+	+/–	+	+/–	+	+
	Forum	–	+	+/–	+	+	+	+	+	+
	Newsgroup	–	+	+/–	+	+	+	+	+	+
	Social Network	–	+	+/–	+/–	+	+	+/–	+/–	+
	Product Review	+	–	–	–	+/–	+	+	–	+
	Wiki	+	–	–	–	–	–	–	–	–
non-UGC	Journalistic Text	+	–	–	–	–	–	–	–	–
	Journalistic Comment	+	–	–	–	–	+	–	–	+
	Encyclopedic Text	+	–	–	–	–	–	–	–	–
	Legal Text	+	–	–	–	–	–	–	–	–
	Technical Literature	+	–	–	–	–	–	–	–	–
	Advertising Text	+	–	–	–	+/–	+/–	+/–	+/–	+

Table 2: Text type classification based on Sandig (1972).

4.2 Text Classification and Support Vector Machines

As described above, regardless of context, humans seem to be able to decide whether a document has been written by a consumer since such texts have inherent attributes that differentiate them from non-UGC. The classification of documents according to inherent attributes relates to the research field of text classification, whose purpose can be described as an assignment problem. Given a set of D docu-

ments and another set of $C = \{c_1, \ldots, c_{|C|}\}$ categories, the task is to assign every pair of $(d_j, c_i) \in D \times C$ one value of T or F. T represents the fact that document d_j can be assigned to class c_i, and F denotes the fact that d_j cannot be assigned to c_i. This classification function $\breve{\Phi} : D \times C \rightarrow \{T, F\}$ is named classifier and aims to approach the unknown objective function $\Phi : D \times C \rightarrow \{T, F\}$, which describes the correct assignment of every document to every class (Sebastiani 2002). However, the correct assignment can often be ambiguous. As seen in the section "Decidability and Attributes of UGC," even with manual classification, different human classifiers may result in different conclusions because assignment is subjective. In manual classification, this divergence is denoted as inter-indexer (in-)consistency (Leonard 1977). Therefore, in text classification, the classifier is often designed in a way that it describes its result in the form of a continuous function within an interval of 0-1 $\widetilde{\Phi_k} : D \times C \rightarrow [0, 1]$.

One of the most popular approaches in text classification is the use of SVMs, which were introduced by Cortes and Vapnik (1995). In benchmarking different classification methods, SVMs usually outperform other approaches, such as Naïve Bayes, decision trees, maximum-entropy, or neuronal nets (Manning et al. 2008; Sebastiani 2002). SVMs are based on the so-called vector space model (Salton et al. 1975). Documents are represented as points in an n-dimensional space. The dimensions are equivalent to n possible terms that might occur within the documents. Each document's vector component represents the weight of a word within the document. This weight can be computed using the binary occurrence; relative frequency; or other metrics, such as the term frequency/inverse document frequency (TF/IDF) (Ramos 2003). If each document is represented as a vector, the similarity (e.g., using the angular distance) of the vectors equals the similarity of the documents the vectors represent. This means that within the vector space model, documents of similar content will be located in close vicinity. This characteristic is used by SVMs when performing certain transformations to separate documents into two classes by a so-called hyperplane. With this hyperplane, every additional document can be classified into one of two classes based on which side of the hyperplane it will be located, which can then be the decision criteria for a binary text classification task.

5 Proposed Approach to UGC Classification

Our approach to demonstrate the feasibility of (semi-)automatic UGC classification centers around the observation that regardless of context, humans are able to decide whether a given document belongs to the category of UGC or non-UGC based on textually inherent signals. We therefore define the task as a binary text classification problem, which we address with supervised machine learning. This distinction is based on the fact that textual UGC incorporates inherent features that are different from non-UGC, as shown above (see "Decidability and Attributes of UGC"). Due to its general performance in text classification, we propose the use of SVMs, which were briefly introduced in the previous section.

Usually, for text classification, not all terms within document collection are used for classification. Rather, a subset, which is denoted as a feature vector, is generally used. The choice of the features has an impact on classification performance. While Silva and Ribeiro (2007) state that selecting features by deleting stopwords (which are usually a language's most frequently used words) enhances accuracy, Joachims (1998) note that in text categorization, only few irrelevant features exist, and feature selection is likely to hurt performance due to information loss.

Therefore, within our UGC-detection feasibility experiment, we test whether filtering out stopwords represented by a language's most frequently used words also removes characteristic signals (e.g., the potentially UGC-indicating personal pronouns "I" or "you"), leading to lower detection accuracy. Because SVMs performance is known not to suffer from a very high number of features (Joachims 1998), it is not problematic to use all distinct words occurring in document collection for the feature vector. Thus, we tokenize every text document into words and only omit those that solely consist of non-word characters, such as ".-#". We use a word's existence as the weight of this word within the feature vector. We denote this feature vector as AT (**A**ll **T**erms). To test if and in what way the removal of stopwords affects classification performance, we compare AT's performance to performance

when applying basic feature selection. As feature-reduction criteria, we utilize DF-thresholding (Yang and Pedersen 1997), which is computationally simple but performs well. In DF-thresholding, terms are excluded from the feature space if the relative occurrence across the documents deceeds a certain threshold. We denote this experiment on feature reduction as DFT.

As DF-thresholding only filters rare terms that are assumed to be uninformative for category prediction and because we are especially interested in those terms that are good at separating categories, we also consider the probability ratio (PR) (Forman 2003) within our second feature selection test which we denote as DFPR (**d**ocument **f**requency, **p**robability **r**atio). PR can be defined as the estimate probability of a term occurring in the positive class (i.e., UGC) divided by the sample estimate probability of the word occurring in the negative class (i.e., non-UGC).

5.1 Data Description and Evaluation

Because a classifier needs a "training set" of examples from which it can learn, we require a collection of documents including annotations as to whether a single document belongs to the class of UGC or non-UGC. To measure the performance of our classification approach, we furthermore need another document collection with the same annotations. This annotated test environment is usually denoted "the gold set" and describes the best result the IR system could achieve (Manning et al. 2008).

With respect to the task of UGC classification, a gold set of English language blog texts including annotations regarding the attribute of "personal context" exists that might be useful for UGC classification (Macdonald et al. 2010). To our knowledge, no similar gold set including the necessary annotations is available for the German language. Therefore, we had to create our own gold set. Within the section "Decidability and Attributes of UGC," we described the Web texts we collected for the synchronization service MobileMe and annotated 500 randomly selected articles. These texts and annotations can now serve as the source of our gold set. During the decidability study, we asked three test subjects to independently classify documents into UGC and non-UGC. Based on these classifications, we perform a majority voting to assign each text to a class. If at least two of the three raters classify a document as belonging to a certain class, it gets assigned to this class within the gold set. As a result, our gold set contains 500 annotated texts, of which 367 texts were classified as UGC, and the remaining 133 as non-UGC. According to the usual practice in text classification and to consistently evaluate the performance of our approach for UGC classification, we utilize 10-fold cross validation for testing our proposed approach and provide the usual metrics of precision, recall, accuracy, and F1-Score, which are defined as follows.

$$tp = true\_positive$$

$$tn = true\_negative$$

$$fp = false\_positive$$

$$fn = false\_negative$$

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$accurracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Equation 1: Evaluation metrics

As described above, within our first experiment denoted as AT, we do not perform any feature reduction and instead use all terms within the document collection as a feature vector. For the SVM classification task, we utilize the library libSVM (Chang and Lin 2011) and follow the suggestions of Hsu et al. (2003) to determine the SVM parameter values. Within the AT experiment, the resulting feature vector contains 9,430 distinct terms. Due to the amount of features, we first test the application of a linear kernel (c = 9) that results to an accuracy of 75% with very high precision for UGC detection but very weak precision for non-UGC (~6.767%). Changing the SVM kernel to a radial basis function kernel (RBF) (using c = 4, gamma = -10) barely increases the overall accuracy to 75.4% and precision for non-UGC classification to 10.526% but lowers recall.

Within the second experiment DFT, we test DF-thresholding as a feature-selection approach and receive an increased overall accuracy of 81.4% (versus 75.4%) while simultaneously increasing recall for UGC/non-UGC detection. Although the precision for UGC classification decreases from 98.91% to 94.823%, non-UGC precision dramatically increases from 10.526% to 44.361%. This also leads to a strong enhancement of F1 for non-UGC, while F1 for UGC also increases from 85.51% to 88.21%.

Finally, within the third experiment DFPR, we combine DF-thresholding (see experiment DFT) and probability ratio (PR) and receive the best overall accuracy, which amounts to 87% for 10-fold cross validation. The best results are achieved on a minimum DF-threshold of 0.068 and a minimum normalized PR of 0.35, which leads to a feature vector of 134 discriminating terms. While precision for UGC detection remains nearly the same (94.823% versus 94.005%), recall increases from 82.464% to 88.918%. We furthermore observe a strong rise in the precision for non-UGC classification to 67.669% (versus 44.361%) as well as an increase in recall to 80.357% (compared to DFT amounting to 75.641%). As a result, DFPR leads to the best overall results both for F1 (UGC) and F1 (non-UGC). The results on our 10-fold cross validation are summarized in Table 3.

	Accuracy	**Precision (UGC)**	**Recall (UGC)**	**Precision (Non-UGC)**	**Recall (Non-UGC)**	**F1 (UGC)**	**F1 (Non-UGC)**
AT (linear)	75.000%	**99.728%**	74.694%	6.767%	**90.000%**	85.41%	12.58%
AT (RBF)	75.400%	98.910%	75.311%	10.526%	77.778%	85.51%	18.54%
DFT (RBF)	81.400%	94.823%	82.464%	44.361%	75.641%	88.21%	55.92%
DFPR (RBF)	**87.000%**	94.005%	**88.918%**	**67.669%**	80.357%	**91.39%**	**73.46%**

Table 3: 10-fold cross-validation results of UGC/non-UGC classification

From the classification results, it can be seen that the semi-supervised classification of UGC using SVMs is basically feasible and that feature selection clearly leads to better results than not restricting the term feature vector.

6 Discussion and Future Research

Within this article, we were able to show that the automatic classification of textual content into UGC versus non-UGC is possible. We had to assume that authors and their motivation(s) (according to the UGC definition) are reflected in the characteristics of the published content. This was generally proven by the overall agreement of different independent annotators deciding whether documents were created by users. Nevertheless, this assumption has to be questioned when authors explicitly use or avoid attributes that are usually found in UGC. For example, this is the case in online encyclopedias, such as Wikipedia, for which the linguistic style is consciously formal and objective. The same applies to texts that have been written by advertising copywriters, who often mimic their customers' informal language in order to better relay their advertising message, as well as for astroturfing, opinion spam, and paid postings, for which commercial organizations and agencies write content explicitly mimicking user conversations to mislead consumers in their quest for information. Still, the sheer existence of

this textual mimicry hints at the fact that UGC signals must exist within text. The detection of these content types has also gained rising attention by the research community in previous years (see Hu et al. 2011; Jindal and Liu 2008; Malbon 2013; Mukherjee et al. 2013; Xu and Zhao 2012). We assume that our approach is not able to detect these kinds of content, but if characteristics of UGC are available, classification according to UGC/non-UGC using traditional techniques from text classification may lead to convincing results. Regarding paid postings and astroturfing, it has to be questioned if this limitation of the UGC-detection approach—namely, not being able to recognize these kinds of content properly—is problematic. As long as human annotators and readers are unable to differentiate paid postings from regular UGC and judging from a perspective regarding the influence of such texts on other readers, it makes little difference if they are in fact UGC. Therefore, it may be justifiable to consider them in voice-of-the-customer analyses since this kind of content also influences other consumers in search of, for example, product reviews.

Although we have basically shown the feasibility of automatic UGC classification there are open issues that need to be addressed in future research. These future research opportunities can be divided into applicability, performance, and evaluation.

Regarding applicability, an extension of our approach to languages other than German would be an interesting research path. We assume that textual content from other languages also incorporates attributes that qualify for automatic UGC detection.

Furthermore, we only tested our approach on a certain domain of software/communication. Other information demands from other domains should be considered in order to confirm that the approach is applicable in general.

Within our study, we focused on the classification of Internet documents, the length of which is usually longer than other sources, such as Twitter tweets (140 characters). As such, future research should explore how the length of text affects classification performance and if small specific texts (e.g., tweets) with different characteristics are also classifiable as UGC/non-UGC.

Although we have shown that our approach to UGC detection already performs well within the chosen domain, our approach should be considered as a baseline demonstrating basic feasibility. The evaluation of other feature-weighting/selection approaches (e.g., TF/IDF, information gain, mutual information, chi-square) as well as other classification techniques (e.g., Naïve Bayes, conditional random fields, or maximum entropy models) should be considered to further increase classification performance.

Regarding evaluation, future research could directly compare context-free human classification performance with system-performance within different domains. Also, the creation of larger gold sets for different languages and domains to ensure the comparability of systems is desirable.

Finally—and from our point of view—the topic of automatically detecting UGC on the Web will be of increasing importance as the Web's entry barriers continue to decrease while user participation and content creation increases. Organizations and academics are interested in silently listening to the online consumers' voice more and more. Nevertheless, the volume and velocity of public available data from the web requires the utilization of automated analysis and (social media) monitoring systems. Those systems should help researchers to listen to the right research subjects (i.e. consumers). Therefore, it is our hope that our work encourages other researchers to contribute to the topic of the article at hand.

References

Bauer, C. A. 2010. User Generated Content–Urheberrechtliche Zulässigkeit nutzergenerierter Medieninhalte, Springer.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. 2011. "Discriminating gender on Twitter," Presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309.

Chang, C.-C., and Lin, C.-J. 2011. "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST) (2:3)ACM, p. 27.

Corney, M., de Vel, O., Anderson, A., and Mohay, G. 2002. "Gender-preferential text mining of e-mail discourse," Presented at the Computer Security Applications Conference, 2002. Proceedings. 18th Annual, pp. 282–289.

Cortes, C., and Vapnik, V. 1995. "Support-vector networks," Mach. Learn. (20:3)Springer, pp. 273–297.

Daugherty, T., Eastin, M. S., and Bright, L. 2008. "Exploring consumer motivations for creating user-generated content," Journal of Interactive Advertising (8:2)The University of Sydney, pp. 1–24.

Decker, R., and Trusov, M. 2010. "Estimating aggregate consumer preferences from online product reviews," International Journal of Research in Marketing (27:4), pp. 293–307.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. 2003. "Authorship Attribution with Support Vector Machines," Applied Intelligence (19:1-2)Kluwer Academic Publishers, pp. 109–123.

Egger, M., and Lang, A. 2013. "A Brief Tutorial on How to Extract Information from User-Generated Content (UGC)," KI-Künstliche Intelligenz (27:1)Springer, pp. 53–60.

Feldkamp, J. 2007. "Qualitätsaspekte im User-Generated Content," Internetökonomie und Hybridität, p. 39.

Forman, G. 2003. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Mach. Learn. Res. (3)JMLR.org, pp. 1289–1305.

Grob, H., and Vossen, G. 2007. Entwicklungen im Web 2.0 aus technischer, ökonomischer und sozialer Sicht, (D. Ahlert, D. Afderheide, K. Backhaus, J. Becker, H. Grob, K. Hartwig, T. Hoeren, H. Holling, B. Holznagel, S. Klein, A. Pfingsten, and K. Röder, eds.), Münster : ERCIS, pp. 1–240.

Hermida, A., and Thurman, N. 2008. "A clash of cultures: The integration of user-generated content within professional journalistic frameworks at British newspaper websites," Journalism Practice (2:3)Taylor & Francis, pp. 343–356.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J., 2003. A Practical Guide to Support Vector Classification, Department of Computer Science, National Taiwan University.

Hu, N., Liu, L., and Sambamurthy, V. 2011. "Fraud detection in online consumer reviews," Decision Support Systems (50:3), pp. 614–626.

Huang, X., and Croft, W. B. 2009. "A unified relevance model for opinion retrieval," Presented at the Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 947–956.

Jindal, N., and Liu, B. 2008. "Opinion spam and analysis," Proceedings of the international conference on Web search and web data mining, pp. 219–230.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features, Springer.

Kaplan, A. M., and Haenlein, M. 2010. "Users of the world, unite! The challenges and opportunities of Social Media," Business Horizons (53:1), pp. 59–68.

Karlgren, J., and Cutting, D. 1994. "Recognizing text genres with simple metrics using discriminant analysis," Presented at the Proceedings of the 15th conference on Computational linguistics-Volume 2, pp. 1071–1075.

Kessler, B., Numberg, G., and Schütze, H. 1997. "Automatic detection of text genre," Presented at the Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and

Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 32–38.
Kohlschütter, C., Fankhauser, P., and Nejdl, W. 2010. "Boilerplate detection using shallow text features," Presented at the Proceedings of the third ACM international conference on Web search and data mining, pp. 441–450.
Koppel, M., Argamon, S., and Shimoni, A. R. 2002. "Automatically categorizing written texts by author gender," Literary and Linguistic Computing (17:4)ALLC, pp. 401–412.
Krumm, J., Davies, N., and Narayanaswami, C. 2008. "User-Generated Content," IEEE Pervasive Computing (7:4)IEEE, pp. 10–11.
Leonard, L. E. 1977. Inter-indexer consistency studies, 1954-1975: a review of the literature and summary of study results, University of Illinois, Graduate School of Library Science.
Macdonald, C., Santos, R. L., Ounis, I., and Soboroff, I. 2010. "Blog track research at TREC," (Vol. 44) Presented at the ACM SIGIR Forum, pp. 58–75.
Malbon, J. 2013. "Taking Fake Online Consumer Reviews Seriously," Journal of Consumer Policy Springer, pp. 1–19.
Manning, C. D., Raghavan, P., and Schütze, H. 2008. Introduction to information retrieval, (Vol. 1) Cambridge University Press Cambridge.
Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. 2013. "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews," UIC-CS-03-2013. Technical Report.
Pang, B., and Lee, L. 2008. "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval (2:1-2)Now Publishers Inc., pp. 1–135.
Park, D.-H., Lee, J., and Han, I. 2007. "The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement," International Journal of Electronic Commerce (11:4)ME Sharpe, pp. 125–148.
Ramos, J. 2003. "Using TF-IDF to Determine Word Relevance in Document Queries," Presented at the Proceedings of the First Instructional Conference on Machine Learning.
Randolph, J. J., Thanks, A., Bednarik, R., and Myller, N. 2005. "Free-marginal multirater kappa (multirater κfree): an alternative to Fleiss' fixed-Marginal multirater kappa," Presented at the Joensuu learning and instruction symposium.
Salton, G., Wong, A., and Yang, C.-S. 1975. "A vector space model for automatic indexing," Communications of the ACM (18:11)ACM, pp. 613–620.
Sandig, B. 1972. "Zur Differenzierung gebrauchssprachlicher Textsorten im Deutschen," Gülich, Elisabeth/Raible, W.(Hgg.) Textsorten. Frankfurt aM: Athenäum, pp. 113–124.
Sebastiani, F. 2002. "Machine learning in automated text categorization," ACM Comput. Surv. (34:1)New York, NY, USA: ACM, pp. 1–47.
Silva, C., and Ribeiro, B. 2007. "On text-based mining with active learning and background knowledge using SVM," Soft Computing (11:6)Springer, pp. 519–530.
Solorio, T., Hasan, R., and Mizan, M. 2013. "A Case Study of Sockpuppet Detection in Wikipedia," NAACL 2013, p. 59.
Stamatatos, E. 2009. "A survey of modern authorship attribution methods," Journal of the American Society for Information Science and Technology (60:3)Wiley Online Library, pp. 538–556.
Stein, D. 2002. "Sprache im Internet - Internet in Universität und Wirtschaft," in Jahrbuch der Heinrich-Heine-Universität Düsseldorf 2002, Düsseldorf: Gert Kaiser, pp. 305–320.
Vater, H. 2001. "Einführung in die Textlinguistik: Struktur und Verstehen von Texten. 3. überarbeitete Aufl," München: Fink Verlag.
Vickery, G., and Wunsch-Vincent, S. 2007. Participative Web And User-Created Content: Web 2.0 Wikis and Social Networking, Paris, , France: Organization for Economic Cooperation and Development (OECD).
Weimer, M., Gurevych, I., and Mühlhäuser, M. 2007. "Automatically assessing the post quality in online discussions on software," Presented at the Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Stroudsburg, PA, USA, pp. 125–128.

Xu, Q., and Zhao, H. 2012. "Using Deep Linguistic Features for Finding Deceptive Opinion Spam.," Presented at the COLING (Posters), pp. 1341–1350.

Yan, X., and Yan, L. 2006. "Gender Classification of Weblog Authors.," Presented at the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 228–230.

Yang, Y., and Pedersen, J. O. 1997. "A comparative study on feature selection in text categorization," (Vol. 97) Presented at the ICML, pp. 412–420.

Zhang, W., Yu, C., and Meng, W. 2007. "Opinion retrieval from blogs," Presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 831–840.