Sentiment analysis of informal textual communication in cyberspace

Georgios Paltoglou[a], Stephane Gobron[b], Marcin Skowron[c], Mike Thelwall[a], and Daniel Thalmann[b]

[a]School of Computing and Information Technology, University of Wolverhampton,
Wulfruna Street, Wolverhampton WV1 1LY, UK
[b]Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[c]Austrian Research Institute for Artificial Intelligence, 1010 Vienna, Austria
g.paltoglou@wlv.ac.uk, stephane.gobron@epfl.ch, marcin.skowron@ofai.at,
m.thelwall@wlv.ac.uk, daniel.thalmann@epfl.ch

Abstract. The ability to correctly identify the existence and polarity of emotion in informal, textual communication is a very important part of a realistic and immersive 3D environment where people communicate with one another through avatars or with an automated system. Such a feature would provide the system the ability to realistically represent the mood and intentions of the participants, thus greatly enhancing their experience. In this paper, we study and compare a number of approaches for detecting whether a textual utterance is of objective or subjective nature and in the latter case detecting the polarity of the utterance (i.e. positive vs. negative). Experiments are carried out on a real corpus of social exchanges in cyberspace and general conclusions are presented.

Keywords: Opinion Mining, Sentiment Analysis, Conversational Systems, Virtual Reality, Virtual Human, Emotional Profile

1 Introduction

The proliferation of social networks such as blogs, forums and other online means of expression and communication have resulted in a landscape where people are able to freely discuss online through a variety of means and applications.

Probably one of the most novel and interesting way of communication in cyberspace is through 3D virtual environments. In such environments, people, represented by their avatars, socialize and interact with each other and with virtual humans operated by machines i.e., computer systems. Examples of such virtual environments are flourishing and include Second Life[1], World of Warcraft [2], There[3], IMVU[4], Moove[5], Activeworlds[6], Bluemars[7], Club Cooee[8], etc.

Despite the fact that the graphics of those environments remain relatively poor, futuristic movies such as Avatar[9] provide an example of sophisticated landscapes and renderings that will be attainable by such environments in the foreseeable future. However, regardless of how attractive and realistic such artificial 3D worlds become, they will always remain heavily dependant on the quality of human communication that takes place within them. As shown in [17, 4, 15], communication in environments that are not limited to one, textual modality, consists of not just semantic data transfer, but also of dense non-verbal communication where sentiment plays an important role. Moreover, without emotion no consistent and coherent (virtual) body language is possible. Such primordial movements include facial expressions, eye looks, arm-language coordination, etc.

Sentiment detection from textual utterances can play an important role in the development of realistic and interactive dialog systems. Such systems serve various educational, business or entertainment oriented functions and also include systems that are deployed in 3D virtual environments. With the aid of "dialog coherence" modules, conversational systems aim at a realistic interaction flow at the emotional level e.g., Affect Listeners [35] and can greatly benefit from the correct identification of the emotional state of their participants. Taking into consideration that the majority of input to practical conversational systems constitute of short, informal, textual exchanges, it is essential that the sentiment analysis component integrated in the dialog system is able to cope with this type of informal, often incomplete or ill-formed type of communication.

Sentiment analysis, the process of automatically detecting if a text segment contains emotional or opinionated content and extracting its polarity or valence, is a field of research that has received significant attention in recent years, both in academia and in industry. The aforementioned increase of user-generated content on the web has resulted in a wealth of information that is potentially of vital importance to institutions and companies, providing them with data to research their consumers, manage their reputations and identify new opportunities. As a result, most of the research in the field has been limited to product reviews (i.e. [12, 42]), where the aim is to predict whether the reviewer recommends a product or not, based on the textual content of the review.

The focus of this paper is different. Instead of focusing our attention to product reviews, we explore a more ubiquitous field of informal, social interactions in cyberspace. The unprecedented popularity of social platforms such as Facebook,

[1] http://secondlife.com
[2] http://www.worldofwarcraft.com
[3] http://www.there.com
[4] http://www.imvu.com
[5] http://www.moove.com
[6] http://www.activeworlds.com
[7] http://bluemars.com
[8] http://www.clubcooee.com
[9] http://www.avatarmovie.com/

Twitter, MySpace as well as 3D virtual worlds has resulted in an unparallel increase of textual exchanges that remains relatively unexplored especially in terms of its emotional content.

Specifically, we aim to answer the following question: can lexicon-based approaches perform more effectively than machine-learning approaches in this domain? This question is particularly important, because previous research in sentiment analysis using product reviews has shown that machine-learning approaches typically outperform lexicon-based ones but no exploration of whether the same holds for informal, social interactions has been carried in the past. The difference between the two domains are numerous. Firstly, reviews tend to be longer and more verbose than typical social interactions which may only be a few words long and often contain significant spelling errors [40]. Secondly, no clear "golden standard" exists in the domain of informal communications with which to train a machine-learning classifier in opposition to the "thumbs up" or "thumbs down" feature of reviews. Lastly, social exchanges on the web tend to be much more diverse in terms of their topics with issues ranging from politics and recent news to religion while in contrast, product reviews by definition have a specific subject, i.e. the product under discussion.

The study of emotional and social interactions in virtual worlds imply the study of virtual human (VH) behaviors. Two types of VH exist: avatars (i.e. the projection of a real human in the 3D environment) and agents (i.e. the projection of an autonomous machine simulating a human in the virtual world). These VH types result in three possible types of communications: avatar to avatar, agent to agent and avatar to agent. Each one of those has the following interesting aspects respectively:

- A non verbal body language based on VH emotional states and mind profile.
- A potential visualization of the interaction from a third VH that should be represented by an avatar;
- A non-verbal communication for the human representation and an action of agent strongly influenced by interpreted emotions from the avatar.

It seems only logical that artificial intelligence and conversation systems would strongly benefit these aspects in order to make the communication more realistic.

The structure of this paper is as follows. The next section provides a brief overview of relevant work in sentiment analysis. Section 3 presents the lexicon-based classifier and section 4 presents the two machine-learning classifiers that will be used in this study. Section 5 describes the data sets that were used and explains the experimental setup while section 6 presents and analyzes the results. Finally, we conclude and present some potential future directions of research.

2 Prior Work

Sentiment analysis, also known as opinion mining, has known considerable interest recently. Most research has focused on analyzing the content of either movie or general product reviews (e.g. [31, 5, 12]). Attempts to expand the application of sentiment analysis to other domains, such as debates [41, 19], news [13] and blogs [27, 24] are also prominent. The seminal book of Pang and Lee [29] presents a thorough analysis of the work in the field. In this section we will focus on the more prominent work which is relevant to our approach.

Pang et al. [31] were amongst of the first to explore the sentiment analysis of reviews, focusing on machine-learning approaches. These approaches generally function as follows: initially, a general inductive process *learns* the characteristics of a class during a training phase, by observing the properties of a number of preclassified documents (i.e. *reference corpus*) and applies the acquired knowledge to determine the best category for new, unseen documents, during testing. Pang et al. [31] experimented with three different algorithms: Support Vector Machines (SVMs), Naive Bayes and Maximum Entropy classifiers, using a variety of features, such as unigrams and bigrams, part-of-speech tags, binary and term frequency feature weights and others. Their best attained accuracy in a dataset consisting of movie reviews, was attained using a SVM classifier with binary features, although all three classifiers gave very comparable performance. Other approaches (e.g. [25, 45, 47]) have focused on extending the feature set with semantically or linguistically-driven features in order to improve classification accuracy.

Dictionary/lexicon-based sentiment analysis is typically based on lists of words with some sort of pre-determined emotional weight. Examples of such dictionaries include the General Inquirer (GI) dictionary [46] and the "Linguistic Inquiry and Word Count" (LIWC) software [33], which are also used in the present study. Both lexicons are build with the aid of "experts" that classify certain tokens in terms of their affective content (e.g. positive or negative). The "Affective Norms for English Words" (ANEW) lexicon [6] contains ratings of terms on a nine-point scale in regard to three individual dimensions: valence, arousal and dominance. The ratings were produced manually by psychology class students. Ways to produce such "emotional" dictionaries in an automatic or semi-automatic fashion have also been introduced in research [43, 7, 37]. Emotional dictionaries have mostly been utilized in psychology or sociology oriented research [10, 36].

The idea of emotional conversationalists is relatively old. First attempts to create such a system can be traced back to Parry [11], a chatterbot intended for studying the nature of paranoia and able to express fears, anxieties or beliefs. More recent work include research on the development of synthetic characters and chatterbots with personalities [2, 14] and studies on emotional responses and their influence on the creation of believable agents or interactive virtual personalities [3, 22]. In [1] authors focused on the role of emotions for gaining rapport in spoken dialog systems by rendering responses that contain suitable emotion, both lexically and auditory. Studies on the role of facial expressions in building rapport in a virtual human-users interactions were conducted in [16]. A chatterbot system that generates emotional responses by selecting and displaying expressive images of the character emulated by the chatterbot was presented in [44].

It has been almost two decades that emotional communication for virtual worlds is a challenging research field. One of the pioneer paper has been proposed by Cassel et al. [9]. In the proposed system, conversations between multiple human-like agents were automatically generates and animates with appropriate and synchronized speech, intonation, facial expressions, and hand gestures. [8] proposed numerous ways to design personality and emotion models for virtual humans. More recently, [39] predicted a specific personality and emotional states from hierarchical fuzzy rules to facilitate personality and emotion control, and in 2009, Pelachaud et al. [32] developed a model of behavior expressivity using a set of six parameters that act as modulation of behavior animation. Finally, this year, [15] introduced a graphical representation of human emotion extracted from text sentences. The main contributions of that approach included an original pipeline that extracts, processes, and renders emotion of 3D VH. Additionally, the paper presented methods to optimize the computational pipeline so that real-time virtual reality rendering can be achieved on common PCs. Lastly, it was demonstrated how the Poisson distribution can be utilized to transfer database extracted lexical and language parameters into coherent intensities of valence and arousal (i.e. parameters of Russell's circumplex model of emotion).

3 Lexicon-based classifier

The lexicon-based classifier is a typical example of an *unsupervised* approach, because it can function without any *reference corpus* and doesn't require any training (i.e. can be applied "off-the-shelf"). In contrast to most previous approaches in opinion mining, the classifier provides not a binary judgement of polarity (i.e. positive or negative) but instead two independent ratings; one for the positive dimension ($C_{pos} = \{+1, +2, +3, +4, +5\}$) and one for the negative ($C_{neg} = \{-1, -2, -3, -4, -5\}$). Higher absolute values indicate stronger emotion and values $\{1, -1\}$ indicate lack of (i.e. objective text).

For example, a score like $\{+3, -1\}$ would indicate the presence of only positive emotion, $\{+1, -4\}$ would indicate the presence of (quite strong) negative emotion and $\{+4, -5\}$ would indicate the presence of both negative and positive emotion. When applied to binary classification, the emotion with the highest absolute value is returned as the final judgement. We solve conflicts of equality (e.g. $\{+3, -3\}$) by taking into consideration the number of positive and negative tokens and giving preference to the class with the largest number of tokens.

The algorithm is based on two, complimentary, *emotional* dictionaries in order to extract the polarity (positive or negative) of terms. The first lexicon is the General Inquirer (GI) lexicon [38], from which we extracted the positive and negative word lists. The GI lexicon has often been used in research as the "golden standard" for algorithms that aim to automatically extract the sentimental orientation of words [43]. The second lexicon is extracted from the "Linguistic Inquiry and Word Count" (LIWC) [33] software[10] which was derived from a

[10] http://www.liwc.net

number of psychological studies and maintains an extensive dictionary list along with human assigned emotional categories and strengths for each lemma.

The classifier works in a simple, rule-based manner; given a document D, the algorithm detects all words that belong to either emotional dictionary and extracts their polarity and intensity. We modify the initial term scores with additional, linguistically-driven functionalities such as: *negation detection* (e.g. “good” versus “not good”), *capitalization detection* (e.g. “bad” versus “BAD”), *exclamation and emoticon detection* (e.g. “happy!!” or “:-)”) *intensifiers* (e.g. “liked” versus “liked very much”) and *diminishers* (e.g. “excellent” versus “rather excellent”), to produce the final document scores [18].

All modules function in the following way: the neighborhood of every word that is present in the text and belongs to either the GI or the LIWC lexicons, is scanned for “special” terms, such as negators (e.g. “not”) intensifiers (e.g. “very”) or diminishers (e.g. “little”). Neighborhood is defined as the area 5 words before and after the emotional term or the end or beginning of the sentence, whichever comes first. The specific span of the neighborhood was chosen after some initial experiments, not reported here because of space constraints. If an intensifier or diminisher word is found then the absolute original emotional value of the word is modified: increased by one in the former case and diminished by one in the latter. For example, if “bad” has an initial value of -3 then ”very bad” would be modified to -4. Similarly, “somewhat good” would be judged as +2, taking into consideration that “good” has an original value of +3.

If a negation term is found then the absolute value of the emotional term is decreased by 1 and its polarity is reversed. For example “not bad” would be +2. The intuition behind the reduction by one (instead of a simpler reversal of signs) is that although the polarity of a term is reversed with the usage of negation, the full original emotional weight of a term (such as “bad” in the above example) isn’t fully transferred to the other class and thus the reduction by one. Simply put, one doesn’t typically use “not bad” if one means “good”.

Lastly, for the *capitalization detection* module, if a word that is written in capital letters only is detected within the neighborhood of an emotional word, then the weight of the word is modified in the same manner as if an *intensifier* was detected. The *exclamation and emoticon detection* module also functions in the same manner.

The score of a document on the C_{pos} and C_{neg} scales is the maximum positive and negative number produced respectively. As previously stated, when the classifier is used for binary positive/negative classification then the class with the highest absolute value is considered dominant. Let it be noted that a document is classified as objective iff its scores are $\{+1, -1\}$.

4 Machine-Learning classifiers

The problem of detecting and extracting the emotional content of text has also been approached from the *machine-learning* perspective. The goal of this approach is to design and develop algorithms that provided with a set of documents with pre-assigned classes (i.e. training set) automatically learn patterns that will help correctly classify new documents.

In this paper, we mainly deal with binary classification problems, where the aim of the classifier is to detect whether a piece of text belongs to one of two mutually exclusive categories. We will apply this setting to two different classification problems; in the first case, the classifier will aim to detect whether a piece of text is objective or subjective and in the second case whether a pre-assigned subjective text contains positive or negative emotion.

We present experiments with two classifiers, both of which as considered state-of-the-art: a Naive Bayes and a Maximum Entropy classifier [23]. Previous experiments have shown that both classifiers perform very well in general classification tasks [26] and more specifically in opinion mining applications [31, 30]. As it is typical for text categorization problems we represent documents using the standard bag-of-words approach, therefore each document D is represented as: $D = \{w_1, w_2, ..., w_m\}$ where w_i is typically a single word (i.e. token) and m is the number of unique words in the training set[11].

Both classifiers function on the same premise, i.e. to maximize the posterior probability $P(c|D)$ that document D belongs in class c. Typically the best class is the *maximum a posteriori* (MAP) class c_{MAP}:

$$c_{MAP} = arg \max_{c \in C} \{P(c|D)\} \tag{1}$$

In practice, that means that we estimate $P(c_i|D)$ for all i's and choose the class with the highest probability. The way that $P(c|D)$ is estimated by the Naive Bayes and the Maximum Entropy classifier differs significantly. In the next two sections we briefly describe both classifiers.

4.1 Naive Bayes classifier

We apply the Bayes rule to equation 1:

$$c_{MAP} = arg \max_{c \in C} \left\{ \frac{P(D|c) * P(c)}{P(D)} \right\} \propto arg \max_{c \in C} \{P(D|c) * P(c)\} \tag{2}$$

where we've removed the denominator $P(D)$ since it doesn't influence the outcome of the classification. $P(c)$ if the prior probability of class c, i.e. the relative frequency of the class and can be defined as $P(c) = |\{D|\text{D's class is c}\}|/|D|$. The Naive Bayes classifier assumes that all features are conditionally independent given class c, therefore:

$$P_{NB}(D|c) = \prod_{i=1}^{m} P(w_i|c) = \prod_{i=1}^{m} \frac{\#(w_i, c)}{\#(w_i)} \tag{3}$$

where $\#(w_i, c)$ is the number of times that token w_i has been encountered in documents of class c in the training data set and $\#(w_i)$ is the number of times

[11] Experiments with combinations of two or three words (referred to as bigrams or trigrams in research) have shown no additional advantage.

that the token has occurred in all documents in the training data set. Lastly, in order to avoid zero probabilities, we apply Laplace add-one smoothing, therefore:

$$P_{NB}(D|c) = \prod_{i=1}^{m} \frac{\#(w_i, c) + 1}{\#(w_i) + m} \tag{4}$$

Despite its simplicity, Naive Bayes classifiers have been used in a number of classification tasks and have been found to perform very adequately in most cases.

4.2 Maximum Entropy classifier

The aim of the Maximum Entropy classifier is to find a model that satisfies all the constraints of the problem which also has maximum entropy. The idea behind this goal is that models with less entropy have added information beyond that in the training set, which are not justified by the empirical evidence. Thus, a maximum entropy model aims to preserve as much uncertainty as possible with the condition that the constraints of the problem (i.e. the training data set) are satisfied [26].

The estimation of $P(D|c)$ for Maximum Entropy classifiers takes the following exponential form:

$$P_{ME}(D|c) = \frac{1}{Z(D)} exp(\sum_{i=1}^{m} \lambda_i f_i(D, c)) \tag{5}$$

where $Z(D)$ is a normalization function, that makes sure that the estimated probabilities are within the $\{0, 1\}$ range and $f_i(D, c)$ is a *feature/class* function that is defined as:

$$f_i(D, c) = \begin{cases} 1, \text{ if } w_i \in D \text{ and } D\text{'s class is c} \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

Lastly, λ_i are the parameters of the model that need to be learned during training. We use the Mallet implementation of Maximum Entropy classifiers[12] which uses the improved iterative scaling (IIS) algorithm for finding the optimal parameter values [26].

5 Experimental Setup

We will use two data sets in order to explore whether lexicon-based or machine-learning approaches are better suited to detect subjectivity and polarity in social textual exchanges on the web.

The first data set is extracted from the BBC Messages Boards[13], where registered users are allowed to start discussions and post comments on existing

[12] http://mallet.cs.umass.edu

[13] http://www.bbc.co.uk/messageboards/

discussions on a variety of topics, ranging from UK/World news to religion. Comments at the site are post-moderated and anything that breaks the "House Rules" is deleted. The data set spans from 2005 to 2009 and contains about 2,5 million comments.

The second data set is extracted from the social news website Digg[14], one of the most popular sites on the web where people share and discuss news and ideas. The site is very loosely administered and therefore any kind of language (including profanity) is allowed. The data set spans the months February-April 2009 and contains about 1,6 million individual comments. The difference in language between the two data sets offers a unique opportunity to explore the the effectiveness of emotional classification is both controlled and unrestrained environments.

A small subset of 1,000 comments was sampled from both data sets and given to 3 human annotators to manually annotate their emotional content on two 5-point scales for positive and negative sentiment: `[no positive emotion or energy] +1,+2,...,+5 [very strong positive emotion]` and `[no negative emotion] -1,-2,...,-5 [very strong negative emotion]`. Both data sets and the annotation process are described in detail in [28] and are freely available.

In this paper, we focus on binary classification (objective vs. subjective and positive vs. negative) so we map the original two-dimensional 5-point scale human annotation to a binary scheme in the following manner:

- All the posts that have been rated by the majority of annotators with scores -1 and +1 are considered "objective".
- All posts that have been rated by the majority of annotators with a positive score equal or higher than +3 and have a negative score of -1 or -2 are considered "positive".
- All posts that have been rated by the majority of annotators with a negative score equal or lower than -3 and have a positive score of +1 or +2 are considered "negative".

We use the union of positive and negative posts as "subjective". Although the above process results in a smaller subset of the original 1,000 posts per data set, the remaining posts are much more definitive of their emotional content and some of the ambiguities of the original annotations are removed. In the experiments that are presented we use this subset as the "gold standard". Table 1 presents some statistics about both data sets. Because of the fact that the resulting data set is highly uneven, we use the average value of the $F1$ measure for both classes to quantify classification quality. The $F1$ for class c is defined as:

$$F1_c = \frac{2P_cR_c}{R_c + P_c} \tag{7}$$

[14] http://www.digg.com

Table 1. Number of documents per class for each data set used.

Data set	Number of Documents			Avg. Words	Total Number
	Neutral	Positive	Negative	per Document	of Documents
BBC	96	41	457	63.77	594
Digg	144	107	221	32.91	472

where P and R are the precision and recall that the classifier attaines for class c respectively, defined as:

$$Precision_c = \frac{tp}{tp + fp}, \quad Recall_c = \frac{tp}{tp + fn} \tag{8}$$

where tp is the number of documents correctly classified as belonging to class c ("true positive"), fp is the number of documents falsely classified as belonging to class c ("false positive") and fn is the number of documents falsely classified as not belonging to class c ("false negative"). The final average $F1$ measure is calculated as $F1 = \frac{1}{|c|} \sum_c F1_c$.

Previous research [34] has shown that machine-learning algorithm need data sets of considerable size in order to perform adequately. Therefore, we trained both the Naive Bayes and the Maximum Entropy classifiers on the BLOGS06 dataset [20, 21]. The dataset is comprised of an uncompressed 148GB crawl of approximately 100,000 blogs and their respective RSS feeds. The dataset has been used for 3 consecutive years by the Text REtrieval Conferences (TREC)[15].

Participants of the conference are provided with the task of finding documents (i.e. blog posts) expressing an opinion about specific entities X, which may be people, companies, films etc. The results are given to human assessors who then judge the content of the posts and assign each one a score: "1" if the document contains relevant, factual information about the entity but no expression of opinion, "2" if the document contains an explicit negative opinion towards the entity and "4" is the document contains an explicit positive opinion towards the entity.

We used the produced assessments from all 3 years of the conference to train our classifiers, resulting in 150 different entity searches and 16,481 documents with a score of "1", 7,930 documents with a score of "2" and 9,968 with a score of "4", which were used as the "gold standard" for training our classifiers. For the objective/subjective classification we used the documents that were given a label of "1" as objective and the union of "2" and "4" as subjective. For the positive/negative classification, we used the documents assigned a label of "2" as negative and "4" as positive.

In order to present a thorough examination of the performance of the machine-learning algorithms, we report results with different probability thresholds. Specifically, for the objective/subjective classification we vary the objectivity threshold ($thres_{obj}$) in the $\{0, 1\}$ range using intervals of 0.005. A document is classified

[15] http://www.trec.nist.gov

as objective iff $P(objective|D) > thres_{obj}$. We also do the same for the positive/negative classification varying the positive threshold ($thres_{pos}$), therefore a document is classified as positive iff $P(positive|D) > thres_{pos}$.

6 Results

The results of the objective/subjective classification task on the BBC and Digg data sets are presented in figure 1 and the results of positive/negative classification are presented in figure 2.

In the former case, the experiments generally demonstrate that the simpler lexicon-based classifier is able to attain very good performance, always higher than the Naive Bayes or Maximum Entropy classifiers, even when they utilize optimized parameter thresholds. The results suggest that the simple lack of "emotional" words is enough to provide a strong indication of objectivity. Although Maximum Entropy classifiers have been found to typically outperform Naive Bayes classifiers (e.g. [31]), the same doesn't hold in the specific setting. The results may be attributed to the over-fitting of the former to the training data set, which is different from the testing data set, thus preventing the produced models to generalize effectively in other settings [26]. On the contrary, the Naive Bayes classifier isn't susceptible to the over-fitting problem producing overall better results, especially with subjectivity threshold values in the $\{0.6, 0.8\}$ range.

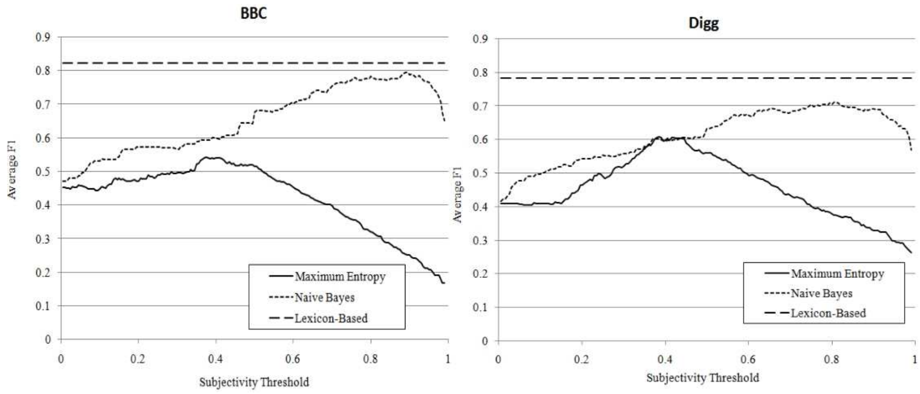

Fig. 1. Average F1 value on the BBC and Digg data sets on the objective/subjective classification task using threshold values for subjectivity in the $\{0, 1\}$ range.

The results from the positive/negative classification (figure 2) task aren't as straight-forward. The lexicon-based classifier isn't able to perform as effectively in this setting especially in the BBC data set, although its performance is much better at the Digg data set. The discrepancy can be attributed to the nature of the discussions on the BBC forums where typically posts tend to be longer[16] and more elaborate, containing both positive and negative words, making it very difficult for the classifier to accurately detect the overall polarity of the post. Digg nonetheless offers a much "easier" setting for the lexicon-based classifier, which attains an average F1 value of 0.74, again much better than even the best-tuned machine-learning classifiers.

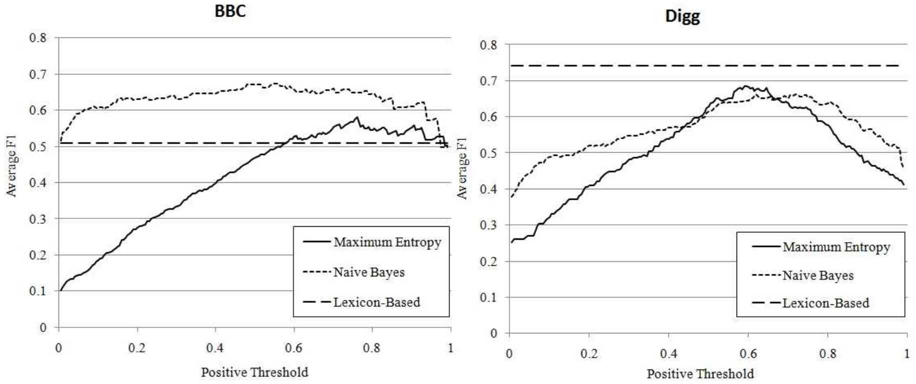

Fig. 2. Average F1 value on the BBC and Digg data sets on the positive/negative classification task using threshold values for positivity in the $\{0, 1\}$ range.

The Naive Bayes classifier again outperforms the Maximum Entropy classifier, strongly indicating that when the training and testing data sets are produced from different word distributions, it provides a more robust solution. Ideally, an intermediate *domain adaptation* procedure between training and testing [5] that would be able to better map the features of the training to the testing set would help increase the effectiveness of both machine-learning approaches, but that is beyond the scope of the paper.

7 Summary and Conclusions

In this paper, we addressed the problem of detecting and analyzing the affective content of textual communication in cyberspace. We argued that the ability to correctly identify the emotional state of the participants of virtual environments based solely on their textual input, is an important part of a realistic and immersive environment that greatly enhances their overall experience.

We approached the problem of sentiment analysis from two different perspectives: supervised, machine-learning approaches and lexicon-based approaches. The former have been used extensively in research, especially in review-oriented

[16] BBC posts have on average two times the number of words in comparison to Digg posts (see table 1).

applications, while the latter, although more intuitive, have had wider dissemination mostly in psychology-based studies.

We presented two state-of-the-art machine-learning algorithms and a lexicon-based classifier which also incorporates a number of linguistically-driven features, such as negation detection, capitalization detection etc. We tested the algorithms in two, diverse (in terms of their content) data sets, both extracted from actual discussions in cyberspace. Individual posts had been annotated from both data sets by three different people in terms of the level of positive or negative content they contained.

The results showed the lexicon-based classifier was able to outperform supervised approaches in the majority of settings, especially in the task of detecting whether textual communication is objective or subjective. On the task of detecting the polarity, the lexicon-based classifier again outperformed other approaches in one of the two data sets, while performing slightly above the baseline in the other. The results indicate a dictionary-based classifier is able to perform adequately in certain environments in cyberspace, elevating the need of developing training corpora for supervised algorithms. Lastly, it was shown that the Naive Bayes classifier was able to offer a more robust performance in comparison to the Maximum Entropy classifier, potentially because of the subtle differences between the training and the testing corpora.

Future research will aim to optimize the term weights of lexicon-based approaches for a given data set, effectively providing the approach with some *supervised* features. Additionally, we aim to further explore the differences between the different media of communication in cyberspace (i.e. fora, im, blogs, virtual worlds) in order to device methods that are able to offer robust performance.

8 Acknowledgements

This work was supported by a European Union grant by the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT. It is part of the CyberEmotions Project (Contract 231323).

References

1. Acosta, J.: Using Emotion to Gain Rapport in a Spoken Dialog System. Ph.D. thesis, University of Texas at El Paso (2009)
2. Barthelemy, F., D.B.G.S., Magnant, X.: Believable synthetic characters in a virtual emarket. In: In Proceedings of the IASTED Artificial Intelligence and Applications (2004)
3. Bates, J.: The role of emotion in believable agents. Communications of the ACM 37(7), 122–125 (1994)
4. Becheiraz, P., Thalmann, D.: A model of nonverbal communication and interpersonal relationship between virtual actors. In: CA '96. p. 58. IEEE Computer Society, Washington, DC, USA (1996)

5. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: 45th ACL. pp. 440–447. Association for Computational Linguistics, Prague, Czech Republic (June 2007)
6. Bradley, M., Lang, P.: Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Tech. rep., Gainesville, FL. The Center for Research in Psychophysiology, University of Florida (1999)
7. Brooke, J., Tofiloski, M., Taboada, M.: Cross-linguistic sentiment analysis: From english to spanish. In: ICRA-NLP (2009)
8. Cassell, J.: Embodied conversational agents. MIT Press, Cambridge, MA, USA (2000)
9. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In: SIGGRAPH '94. pp. 413–420. ACM, New York, NY, USA (1994)
10. Chung, C., Pennebaker, J.: The psychological function of function words. Social communication: Frontiers of social psychology pp. 343–359 (2007)
11. Colby, K.: Artificial paranoia. Artificial Intelligence 2(1), 1–25 (1971)
12. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: WWW '03. pp. 519–528. ACM Press (2003)
13. Devitt, A., Ahmad, K.: Sentiment polarity identification in financial news: A cohesion-based approach. In: Proceedings of the 45th ACL. pp. 984–991. Association for Computational Linguistics, Prague, Czech Republic (June 2007)
14. Galvo, A., B.F.N.A., G., R.:
15. Gobron, S., Ahn, J., Paltoglou, G., Thelwall, M., Thalmann, D.: From sentence to emotion: a real-time three-dimensional graphics metaphor of emotions extracted from text 26(6-8), 505–519 (June 2010)
16. Gratch, J., W.N.G.J.F.E., Duffy, R.:
17. Kappas, A., Hess, U., Scherer, K.R.: Voice and emotion. In: Fundamentals of nonverbal behavior. p. 200238. Cambridge University Press, Cambridge and New York (1991)
18. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence 22(2), 110–125 (2006)
19. Lin, W.H., Wilson, T., Wiebe, J., Hauptmann, A.: Which side are you on? identifying perspectives at the document and sentence levels. In: Proceedings of CoNLL (2006)
20. Macdonald, C., Ounis, I.: The trec blogs06 collection : Creating and analysing a blog test collection. Tech. rep., Department of Computing Science, University of Glasgow (2006)
21. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the trec-2008 blog track. In: TREC 2008 (2008)
22. Magnenat-Thalmann, N.: Creating a smart virtual personality. Lecture Notes in Computer Science 2773(2), 15–16 (1993)
23. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA (1999)
24. Mishne, G.: Experiments with mood classification in blog posts. In: 1st Workshop on Stylistic Analysis Of Text For Information Access (2005)
25. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP

2004. pp. 412–418. Association for Computational Linguistics, Barcelona, Spain (July 2004)

26. Nigam, K., Lafferty, J., Mccallum, A.: Using maximum entropy for text classification. In: IJCAI-99 Workshop on Machine Learning for Information Filtering. pp. 61–67 (1999)
27. Ounis, I., Macdonald, C., Soboroff, I.: Overview of the trec-2008 blog trac. In: The TREC 2008 Proceedings. NIST (2008)
28. Paltoglou, G., Thelwall, M., Buckely, K.: Online textual communcation annotated with grades of emotion strength. In: Proceedings of the Third International Workshop on EMOTION (satellite of LREC): Corpora for research on emotion and affect. pp. 25–31 (2010)
29. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publishers Inc. (2008)
30. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: ACL'04. pp. 271–278 (2004)
31. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: EMNLP 2002 (2002)
32. Pelachaud, C.: Studies on gesture expressivity for a virtual agent. Speech Commun. 51(7), 630–639 (2009)
33. Pennebaker J., F.M., R., B.: Linguistic Inquiry and Word Count: LIWC. Erlbaum Publishers, 2 edn. (2001)
34. Prabowo, R., Thelwall, M.: Sentiment analysis: A combined approach. Journal of Informetrics 3(2), 143 – 157 (2009)
35. Skowron, M.: Affect listeners: Acquisition of affective states by means of conversational systems. In: COST 2102 Training School. pp. 169–181 (2009)
36. Slatcher, R., Chung, C., Pennebaker, J., Stone, L.: Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. Journal of Research in Personality 41(1), 63–75 (2007)
37. Stefano Baccianella, A.E., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: LREC'10 (may 2010)
38. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press (1966)
39. Su, W.P., Pham, B., Wardhani, A.: Personality and emotion-based high-level control of affective story characters. IEEE Transactions on Visualization and Computer Graphics 13(2), 281–293 (2007)
40. Thelwall, M., Wilkinson, D.: Public dialogs in social network sites: What is their purpose? J. Am. Soc. Inf. Sci. Technol. 61(2), 392–404 (2010)
41. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. CoRR abs/cs/0607062 (2006)
42. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: ACL. pp. 417–424 (2002)
43. Turney, P.D., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus. CoRR cs.LG/0212012 (2002)
44. Vrajitoru, D.:
45. Whitelaw, C., Garg, N., Argamon, S.: Using appraisal groups for sentiment analysis. In: CIKM '05. pp. 625–631. ACM, New York, NY, USA (2005)
46. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT/EMNLP 2005. Vancouver, CA (2005)
47. Zaidan, O., Eisner, J., Piatko, C.: Using Annotator Rationales to Improve Machine Learning for Text Categorization. Proceedings of NAACL HLT pp. 260–267 (2007)