What is the Meaning of 5 *'s? An Investigation of the Expression and Rating of Sentiment

Daniel Hardt
Copenhagen Business School
dh.itm@cbs.dk

Julie Wulff
Copenhagen Business School
cbs@juliewulff.dk

Abstract

Do user populations differ systematically in the way they express and rate sentiment? We use large collections of Danish and U.S. film reviews to investigate this question, and we find evidence of important systematic differences: first, positive ratings are far more common in the U.S. data than in the Danish data. Second, highly positive terms occur far more frequently in the U.S. data. Finally, Danish reviewers tend to under-rate their own positive reviews compared to U.S. reviewers. This has potentially far-reaching implications for the interpretation of user ratings, the use of which has exploded in recent years.

1 Introduction

There is a persistent stereotype concerning the way sentiment is expressed and evaluated by Scandinavians and Americans, which is illustrated by these two anecdotes. In the first anecdote, a U.S. researcher gives a talk in a Scandinavian country. After the talk, the researcher is approached by an audience member, who says, "the talk was ok". The U.S. researcher is puzzled by this, until another member of the audience explains to him that this was actually intended to express high praise. The second anecdote: a student at the beginning of his graduate studies at a U.S. university has several meetings with a prominent faculty member, and is repeatedly told that his research ideas are "wonderful". The student is gratified by this, until he overhears other students talking about how this faculty member seems to always respond to ideas by calling them "wonderful".

There is abundant anecdotal evidence that Scandinavians and Americans differ in the way they express and evaluate sentiment: compared to Americans, it seems that Scandinavians downgrade their positive expressions of sentiment. But is this stereotype actually true? In this paper, we investigate this question by analyzing large collections of Danish and U.S. film reviews. These reviews are short pieces of text, combined with a numerical rating which expresses the user's overall evaluation. In our view, such data should provide a meaningful test of the stereotype – if Scandinavians and Americans do indeed differ as we have described, this should be reflected in distributional differences in these datasets.

In particular, the hypothesis concerns distributions of very positive evaluations: compared to U.S. reviewers, we expect a Danish tendency to "downgrade" from very positive to somewhat less positive. We will examine this hypothesis from three different perspectives, in looking at the Danish data vs. the U.S. data:

1. **Ratings:** are there relatively fewer high ratings?

2. **Text:** are there relatively fewer highly positive terms?

3. **Ratings vs. Text:** are there fewer high ratings for texts of a given positivity?

In what follows, we begin with a description of the data sets. Next we examine the distribution of ratings. Then we look at the text positivity: we develop a metric for positivity of terms, and examine their relative distributions. This is followed by an examination of the relation between ratings and texts in the two data sets. We show that the hypothesis is strongly confirmed in all three of its variants. Finally, we observe that these results could have far-reaching implications for the interpretation of recommender systems and user ratings, the use of which has exploded in recent years.

2 Data

The Danish data was downloaded from the Danish movie website scope.dk and contains rated user reviews from 829 films and has a total size of 1,624,049 words. The U.S. data was downloaded from The Internet Movie Database (imdb.com) and contains rated user reviews from 678 films and has a total size of 34,599,486 words.

A search function on www.imdb.com was used to create a list of films and matching IMDb ID tags for films produced in the years 1920-2011. 678 films on the list had a match in the Scope data on title and production year . The IMDb ID tags was used to find the page containing data for each of the films and all reviews which had a correlated rating were downloaded for those 678 films. The U.S. IMDb reviews are rated on a scale of 1 to 10, while the Danish Scope reviews are rated on a scale of 1 to 6.

3 Ratings

Figure 1 gives the number of reviews in each category for *IMDb*.

For IMDb, the top category of 10 has by far the most reviews. For the most part the number of reviews decreases from category 10, with a modest increase in the number of reviews for the lowest category, 1. This distribution makes intuitive sense – it's not surprising that people would be most motivated to write reviews of films they are most enthusiastic about, and, to a lesser extent, also be motivated in cases where they have strong negative feelings. This has been noted in the literature: (Wu and Huberman, 2010) point out that the so-called "brag and moan" view of ratings is fairly typical (as also mentioned by (Hu et al., 2006; Dellarocas and Narayan, 2006)). The tendency of the top category to be the most frequent is also mentioned on the yelp.com site, where the top category of 5 is the most frequent: "The numbers don't lie: people love to talk about the things they love!" (FAQ, 2012).

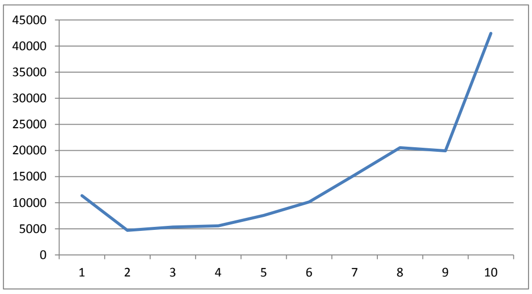

Figure 1: IMDb reviews per category

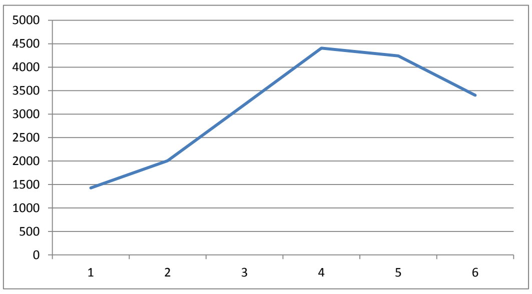

Figure 2: Scope reviews per category

There is a very different distribution in the Danish Scope data, as shown in Figure 2. Here, category 4 (out of 6) is the most frequent. This supports the general prediction that highly positive evaluations are over-represented in the U.S. data compared to the Danish data.

4 Text

We turn now to a second version of our hypothesis: that highly positive terms are over-represented in the U.S. data. We consider highly positive terms to be those that tend to occur in the most positive category and tend not to occur in the other categories.

For each category, we follow (Constant et al., 2009) in defining what they call a *log-odds* distribution for each term, as follows:

$$log\text{-}odds(x_n, R) = ln(\frac{count(x_n,R)}{count(n,R) - count(x_n,R)})$$

Here, n is 1, 2 or 3, denoting terms consisting of one, two or three words (i.e., unigrams, bigrams and trigrams). R is a rating category (1-6 in Scope and 1-10 for IMDb). $Count_n(R)$ is the number of occurrences of all ngrams of length n in Category R, while $count(x_n, R)$ is the number of occurrences of a particular ngram x_n in Category R. Thus we take the log of the number of occurrences of a given ngram in a category, divided by the number of occurrences of all other ngrams in that category.

Intuitively, highly positive terms are those most frequent in the top category and most infrequent in the other categories. Thus we determine positivity as follows:

$$positivity(x_n) = log\text{-}odds(x_n, Rpos) - log\text{-}odds(x_n, Rother)$$

For Scope, *Rpos* is category 6, and *Rother* is categories 1 through 5, while for IMDb *Rpos* is categories 9 and 10, and *Rother* is 1 through 8.

Negativity of terms is defined in a symmetrical fashion:

$$negativity(x_n) = log\text{-}odds(x_n, Rneg) - log\text{-}odds(x_n, Rother)$$

Here, *Rneg* is 1 for Scope and 1 and 2 for IMDb, while *Rother* is 2 through 6 for Scope and 3 through 10 for IMDb.

Tables 1 through 4 give the top 25 most negative and positive terms for both IMDb and Scope. For the negative terms, the most negative terms are at the top of the list, while for the positive terms, the most positive are at the bottom.

Our point of departure is that all terms with positivity greater than 0 are positive terms, while those with negativity less than 0 are negative terms. This gives the ratios of positive to negative terms as shown in Table 5.

Negativity	Term
-5.579750143176	absolutely the worst
-5.47055003096302	the worst piece
-5.47055003096302	or money on
-5.38977979264451	10 worst
-5.30349485263692	money back !
-5.20818412600157	awful movie !
-5.10282306351303	absolutely no redeeming
-5.04493752542859	of worst
-4.98431669202047	! complete
-4.88660293269565	worst piece of
-4.88587585595205	worst piece
-4.85150754157158	horrible waste of
-4.85150754157158	. * from
-4.85150754157158	no redeeming features
-4.85150754157158	the worse movies
-4.85078074773538	... avoid
-4.85078074773538	beyond bad
-4.77740634497507	this is awful
-4.77740282048268	horrible film .
-4.77739929600291	i wasted on
-4.77739929600291	this horrible film
-4.77739929600291	piece of c
-4.6973563149145	what a pile
-4.6973563149145	misfortune of seeing
-4.6973563149145	utter crap </s>

Table 1: 25 most negative terms IMDb

There are somewhat more positive than negative terms in IMDb, and slightly more negative

Positivity	Term
3.66887212985741	gets better every
3.68657177169013	movie . 10
3.70287245596558	sterling hayden
3.70396174809963	top ten movies
3.70396174809963	direction is flawless
3.70396174809963	. outstanding !
3.73786336450852	film . 9
3.73786336450852	masterpiece of film
3.73786336450852	best gangster movie
3.77065325206472	see movie !
3.80131270948848	. greatest
3.80240201511252	... 10 /
3.80240809083371	this masterpiece .
3.83317373851249	. (9
3.8630267663954	movie changed my
3.89092504888785	. 9.5
3.89201436800188	. a 10
3.92751010266618	. 10 /
3.94759374427238	. 10 out
4.02554608429261	+ + </s>
4.07433637792856	favorite movies !
4.20381518291327	! 10
4.41420530401696	! ! 10
4.43932293273085	! 10 /
4.5851638142275	outstanding ! </s>

Table 2: 25 most positive terms IMDb

Negativity	Term
-4.870965702	elendig ! *(terrible)*
-3.867670635	ret elendig *(really terrible)*
-3.666983304	min tid *(my time)*
-3.666958989	noget bras *(some junk)*
-3.577451105	skodfilm *(trash film)*
-3.549191951	ringe ! *(bad)*
-3.531008767	lorte *(crap)*
-3.484669428	elendig </s> *(terrible)*
-3.484669428	ligegyldig film *(meaningless film)*
-3.418380646	ikke engang kan *(can't even)*
-3.415652241	skod </s> *(trash)*
-3.398851791	bras *(junk)*
-3.356843736	elendig film *(terrible film)*
-3.264221321	<s> anonym kedelig *(anon. boring)*
-3.261493243	anonym kedelig *(anon. boring)*
-3.163755010	spilde *(waste)*
-3.149240635	stinker *(stinks)*
-3.141251329	crap
-3.112599209	elendigt *(terrible)*
-3.076666572	uudholdelig *(unbearable)*
-3.038365052	blandt min *(among my)*
-3.030337065	skod *(junk)*
-2.973857889	en elendig *(a terrible)*
-2.973834211	ret nej *(really no)*
-2.942324421	elendig *(terrible)*

Table 3: 25 most negative terms Scope

Positivity	Term
2.87692577130	elsker den *(love it)*
2.87848113547	film den er *(film it is)*
2.88763018980	fantastisk ! </s> *(fantastic !)*
2.89096615230	fantastisk film ! *(fantastic film !)*
2.92051716735	mest geniale *(most genius)*
2.92728613305	kan se igen *(can see again)*
2.95568635350	ret kanon *(really great)*
2.98294109871	jeg elsker den *(i love it)*
3.00279792930	genial </s> *(genius)*
3.02076226510	bedste film jeg *(best film i)*
3.05406651227	mega god *(mega good)*
3.06084014079	6 stjerner . *(6 stars)*
3.11470908673	<s> 6
3.28394697268	bedste film der *(best films that)*
3.40913662108	bedste film nogensinde *(best films ever)*
3.45951064085	geniale film *(genius)*
3.45951064085	film overhovedet *(films at all)*
3.61366505188	fortjener 6 *(deserves 6)*
3.62043477181	ret fantastisk ! *(really fantastic)*
3.75397394578	fed ! ! *(great)*
3.75397394578	ret den bedste *(really the best)*
3.86498694263	simpelthen fantastisk *(simply fantastic)*
3.97713305996	elsker den film *(love the film)*
4.06566510061	6 /
4.94095107482	6 / 6

Table 4: 25 most positive terms Scope

	Positive Terms	**Negative Terms**	**Ratio**
IMDb	50,304,859	46,642,846	1.0785
Scope	1,017,939	1,027,940	0.9903

Table 5: Ratio of positive to negative terms

terms than positive in Scope. However, it is not clear if such a comparison is meaningful. Furthermore, our hypothesis does not concern the total positivity of terms in Danish vs. English, but rather, a difference in the distribution of terms in the most positive categories. To focus our investigation on this issue, we define thresholds very close to zero such that the ratio of positive to negative terms in both data sets is 1.0.

We now can measure the number of occurrences of positive occurrences in each category. As discussed above, our hypothesis is that there should be a difference in distribution of positive terms, especially in the most positive categories. Figures 3 and 4 show that there is indeed a striking difference in distribution.

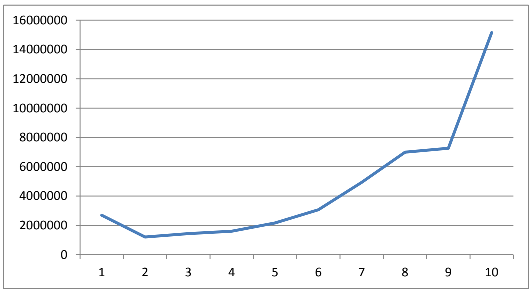

Figure 3: IMDb positive terms per category

5 Ratings vs. Text

We have shown that the hypothesis has been confirmed in two ways: first, there are proportionately more top rated reviews in the U.S. data compared to the Danish data. Second, there are proportionately more occurrences of positive terms in the top categories in the U.S. data vs. the Danish data. We now wish to tease apart these two factors, and pose the question: does the numerical

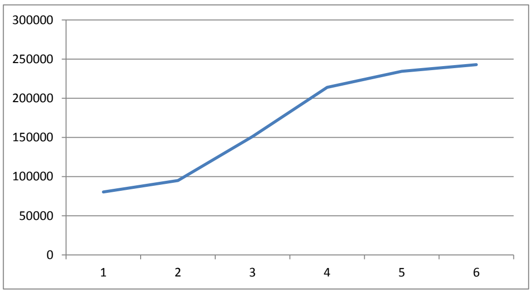

Figure 4: Scope positive terms per category

rating correspond to the positivity of the review?

We define the positivity of a text as the ratio of positive occurrences to negative occurrences in that text. This can be used to assess the positivity of a given review, or the positivity of the complete collection of reviews in a given category. Figures 6 and 5 show the positivity of reviews in each rating category, for Scope and IMDb.

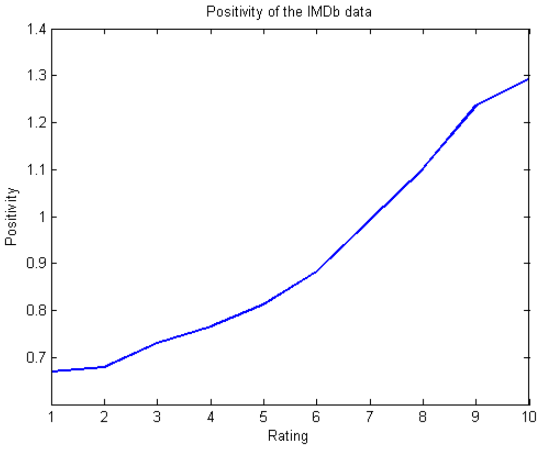

Figure 5: IMDb positivity

Our interest is in the increase in positivity in the highest categories: in IMDb this increase is relatively modest, while it is quite steep for Scope. To assess this difference, we compare the average increase in positivity per category both before and after a category of interest. For Scope, the category of interest is 4: the hypothesis is that reviewers would tend to resist giving ratings higher than 4, even in the face of very positive review text. This is indeed what we find: the rate of change per category above 4 nearly doubles from .16 to .28. We perform a similar analysis with the IMDb data, selecting 8 as the category of interest. Here we find a striking contrast: the rate of change actually drops above 8 (see Table 6).

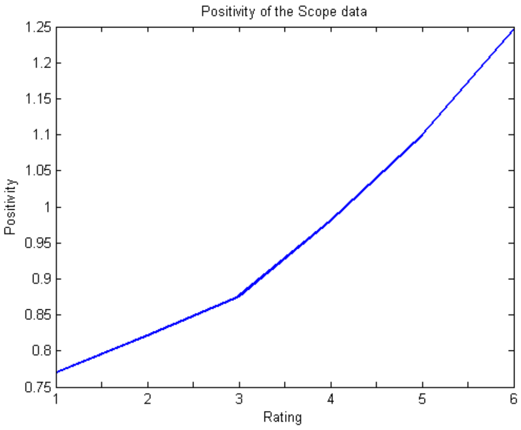

Figure 6: Scope positivity

	Category of Interest	Rate Below	Rate Above	Ratio
Scope	4	.16	.28	.57
IMDb	8	.22	.19	1.15

Table 6: Positivity - Rate of Increase

This analysis strongly supports the third version of our hypothesis: the difference in positivity of U.S. and Danish reviews reflects a difference in the relation of text positivity to rating, for very positive texts. For such texts, Danish reviewers, when compared to U.S. reviewers, have a tendency to "downgrade" a text of a given positivity.

6 Conclusion

There is a widely-held belief that Americans and Scandinavians differ in the way they express and rate positive sentiment. To our knowledge this paper represents the first attempt to test such a belief in a systematic way. Using large collections of film reviews, we have found strong confirmation of the hypothesized difference, defined from three different points of view: ratings, text, and text-rating relations.

In recent years, the use of rating systems have exploded, to the point where they are relied on every day for millions of decisions about everything from where to eat to what film to see, or where and how to take a vacation. The present work, while limited in Scope, suggests a potentially far-reaching conclusion; namely, it points to the possibility that there are systematic differences in rating systems, that we ignore at our peril. As we have seen, Danes differ sharply from Americans in the positivity of ratings and text: they give far fewer top ratings; and the frequency of highly positive terms in the top categories is quite a bit less. One natural conclusion is that there are cultural differences leading Danes to produce reviews and ratings in a rather different way than Americans. In our experience, those familiar with Danish and American culture find this quite plausible and readily suggest numerous potential explanations – perhaps the most compelling of which concerns the traditional grading system in Danish schools[1], where the top grade of "13" was given in only the most exceptional of circumstances, and was always far less frequent than the top grade of "A" in U.S. schools.

There is an obvious alternative explanation for these differences, namely, that Danes are simply less enthusiastic about the films they see. This might seem somewhat paradoxical – since Danes and Americans are both free to choose which films they see, one might expect that they are equally enthusiastic about the films they choose to see and review. However, it has often been suggested that the film industry in many European countries is subject to U.S. cultural imperialism, which would hold that, because of its economic and cultural power, the U.S. film industry is able to substantially alter the film-going options of the Danish public.

We don't discount the possibility that our data in part reflects a general lack of enthusiasm for the films on offer in Denmark, either due to U.S. cultural dominance or perhaps some other factors. This explanation would be rather uninteresting in terms of the general issues concerning the rating and expression of sentiment in different populations, although it ought to be of interest to the producers and distributors of film in Denmark. In any case, we are convinced this is not the complete explanation, because of our third finding, concerning the relation of ratings to text. This shows that there are systematic differences between Danes and Americans for texts expressing a similar level of positivity – Danes tend to move many of these from a top category to a less positive one. In our view this constitutes clear evidence of a systematic difference in how sentiment is treated in the two populations.

We have argued that these differences point to a potentially important problem with the use of rating systems, especially if such differences are widespread. In future work, we intend to examine reviews in other domains, to see if the difference we have found is limited to certain domains or is one that is generally found when comparing Danes and Americans. We are also exploring ways to address the problem these differences pose: one natural hypothesis is that, when there is a systematic mismatch between text and rating, the text positivity is a better guide to the true sentiment. We would like to see if an automatic sentiment analysis might reduce systematic mismatches in these cases.

Acknowledgements

Thanks to Bernardo Huberman for suggesting this line of work, and for lots of feedback and creative input. Chris Zimmerman also gave important feedback and suggestions.

References

Noah Constant, Christopher Davis, Christopher Potts, and Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung*, 1(2):5–21.

C. Dellarocas and R. Narayan. 2006. What motivates consumers to review a product online? a study of the product-specifc antecedents of online movie reviews. In *In Proceedings of the International Conference on Web Information Systems Engineering*.

FAQ. 2012. Yelp.com. `http://www.yelp.com/faq`.

N. Hu, P. A. Pavlou, and J. Zhang. 2006. Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online

[1] The Danish grading system was revised in 2006, in part to make it more in line with grading systems in other countries.(Wikipedia, 2012)

word-of-mouth communication. In *In Proceedings of the ACM Conference on Electronic Commerce*.

Wikipedia. 2012. Academic grading in Denmark — Wikipedia, the free encyclopedia. [Online; accessed 27-August-2012].

Fang Wu and Bernardo Huberman. 2010. Opinion formation under costly expression. *ACM Transactions on Intelligent Systems and Technology*, 1(1).