Stop Words Are Not "Nothing": German Modal Particles and Public Engagement in Social Media

Fabian Rüsenberg[1(✉)], Andrew J. Hampton[2], Valerie L. Shalin[3], and Markus A. Feufel[1]

[1] Department of Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany
fabian.ruesenberg@campus.tu-berlin.de
[2] Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA
[3] Department of Psychology and Kno.e.sis, Wright State University, Dayton, OH 45435, USA
valerie@knoesis.org

Abstract. Social media research often exploits metrics based on frequency counts, e.g., to determine corpus sentiment. Hampton and Shalin [1] introduced an alternative metric examining the style and structure of social media relative to an Internet language baseline. They demonstrated statistically significant differences in lexical choice from tweets collected in a disaster setting relative to the standard. One explanation of this finding is that the Twitter platform, irrespective of disaster setting, and/or specifics of the English language, is responsible for the observed differences . In this paper, we apply the same metric to German corpora, to compare an event-based (the recent election) with a "nothing" crawl, with respect to the use of German modal particles. German modal particles are often used in spoken language and typically regarded as stop words in text mining. This word class is likely to reflect public engagement because of its properties, such as indicating common ground, or reference to previous utterances (i.e. anaphora) [2, 3]. We demonstrate a positive deviation of most modal particles for all corpora relative to general Internet language, consistent with the view that Twitter constitutes a form of conversation. However, the use of modal particles also generally increased in the three corpora related to the 2017 German election relative to the "nothing" corpus. This indicates topic influence beyond platform affordances and supports an interpretation of the German election data as an engaged, collective narrative response to events. Using commonly eliminated features, our finding supports and extends Hampton and Shalin's analysis that relied on pre-selected antonyms and suggests an alternative method to frequency counts to identify corpora that differ in public engagement.

Keywords: Big data · Text mining · Common ground · Collective narrative

1 Introduction

Social media research has demonstrated the capacity to assist in the identification of public response to products and entertainment [4], disaster [5, 6], social phenomena such as gender-based-violence [7] and political events such as elections [8]. Most of these efforts exploit metrics based on frequency counts, sometimes scaled with respect to the

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 89–96, 2018.
https://doi.org/10.1007/978-3-319-93372-6_11

corpus in question. For example, a sentiment analysis might compare the number of positive sentiment words to the number of negative sentiment words in a corpus, to determine net sentiment [9, 10]. Machine learning techniques assist in the identification of diagnostic items [11]. However, a non-existing reference distribution limits interpretation.

1.1 Introducing Relative Metrics

Hampton and Shalin [1] introduced an alternative metric, where observed frequency counts are scaled according to a population-based standard. Exploiting pre-selected lexical items corresponding to physical properties such as size and numerosity, and social properties such as cooperation, Hampton and Shalin demonstrated changes in lexical choice in social media collected during a disaster setting, relative to an Internet language baseline. Such a metric contributes to the development of social alarms, potentially assisting in the management of finite disaster response resources. Similar to sentiment analysis, their metric aims to be domain independent, but focuses more on style than content. Unlike sentiment analysis, and of particular relevance to the present paper, Hampton and Shalin included so-called stop words in their analysis, that is, words that are typically excluded in text mining due to their high prevalence. Consistent with Purohit et al.'s [5] claim that social media behave like a group conversation, Hampton and Shalin interpreted the departures from baseline word patterns as sentinels of breach, emergent from a corpus of tweets and constituting a collective narrative.

Hampton and Shalin's interpretation is not without criticism. In particular, it may be that social media messages are fundamentally different from other forms of Internet material, as Purohit et al.'s conversation analysis indicates. If so, all social media corpora would differ from a broadly composed standard. Moreover, the Hampton and Shalin study was confined to an analysis of English corpora, with its specific linguistic and cultural properties. This paper addresses some of these criticisms. It uses metrics inspired by Hampton and Shalin, applied to German social media corpora drawn from events related to the recent election in Germany, compared to a *Seinfeld*-ian "nothing" corpora concerning Maslow type needs about water, food, shelter, and sleep that were unassociated with any particular event. Continuing in the effort to develop domain-independent metrics, most of the words we examine are common stop words. Continuing in the effort to assess public response to an event, we examine engagement in the recent German elections relative to a "nothing" crawl without an event indicator, by analyzing tweet style and structure rather than its content.

1.2 Modal Particles

The German language contains a unique word class, known as modal particles. These are uninflected words characteristic of spoken language. We choose this particular word class because of its potential to reflect public engagement, indicating the speaker's attitude, referring to common ground, assumptions and expectations of the speaker or receiver, or referencing previous utterances (anaphora) [2, 3]. Linguists identify a core class of 15 modal particles [3, 12–15]: *aber, auch, bloß, denn, doch, eben, eigentlich,*

etwa, halt, ja, mal, nur, schon, vielleicht, wohl. According to [16], all of these words except for *halt* and *mal* are unanalyzed stop words in conventional social media analysis. Examples for the use of modal particles appear in Table 1.

Table 1. Examples of the potential occurrence of a modal particle in a tweet.

Modal particle	Examples
denn/doch	Aber wer soll **denn** in den BT [Bundestag] einziehen? Die FDP besteht **doch** nur aus Lindner, oder? *But who should **"then"** move into the BT (Bundestag)? The FDP consists **"just"** only of Lindner, no?*
nur	Wie kann man **nur** AfD wählen? *How can one **"just"** vote for the AfD?*
vielleicht	Klar, aber wir sollten uns **vielleicht** mit Dingen beschäftigen, die bei uns passieren *Sure, but we should **"maybe"** care about things that happen here.*

German has another welcome property. Relative to English, the use of German is largely confined to Germany and nearby Austria and Switzerland whose citizens are presumably less engaged in the German election. This suspends the need to rely on location metadata for message source and maximizes location-specific data collection.

We demonstrate that the prevalence of modal particle increases for an event of national significance relative to "nothing", suggesting public engagement. We use the method of Hampton and Shalin as described in Sect. 2.1.

2 Methods

2.1 Dataset

Several thousand unique tweets in the German language were collected for three events related to the German Elections 2017 relative to a "nothing" crawl referring to basic human needs (see Sect. 1.1). Table 2 gives an overview of the events, the time frame in which data were collected and the keywords used. Relative to the event corpora, which by definition include election keywords, tweets in the "nothing" corpus do not contain election keywords. Tweets were obtained either through an online semantic web application, Twitris [17], or via the Twitter Search API in R using RStudio (Version 1.0.136). From the original data set, unique tweets were obtained by eliminating retweets and using the `unique()` command in R.

Table 2. Overview of events including start and end date and crawling word set.

Event	Start	End	Crawling word set
German Elections 1 $N = 601{,}498$	2017 Sep 20	2017 Sep 26	spd, cdu, fdp, afd, npd, grüne, gruene, linke, union, #btw17, #btw2017, bundestagswahl, bundestag, wahlen, deutschland, land, partei, merkel, schulz, stimme, demokratie, wahlkampf, hochrechnung
German Elections 2 $N = 183{,}861$	2017 Oct 11	2017 Oct 23	spd, cdu, fdp, afd, npd, grüne, gruene, linke, #btw17, #btw2017, bundestagswahl, bundestag, wahlen, partei, stimme, groko, demokratie, wahlkampf, hochrechnung, #groko, jamaika, #jamaika, wahlergebnis
German Elections 3 $N = 85{,}689$	2018 Mar 02	2018 Mar 06	spd, cdu, csu, union, groko, #groko, grogo, #grogo, #nogroko, #spderneuern, merkel, neuwahlen, regierung, koalition, koalitionsvertrag, minderheitsregierung, jamaika, mitgliedervotum, bundestagswahl
Nothing $N = 61{,}831$	2018 Mar 01	2018 Mar 06	wasser, getränk, trinken, frühstück, mittagessen, abendessen, brunch, snack, essen, haus, wohnung, appartment, schlaf, schlafen

Word Frequency Norms. Baseline frequencies of words in German Internet language can be found in a collection of linguistically processed web corpora called DECOW [18, 19]. In order to standardize comparison between pairs with different absolute frequencies, the following approach was used, exploiting the inability to interlink modal particles, i.e., to combine them by using the German words *und* and *oder* (and, or) [2] (see Eq. 1). We compare the observed proportion in the data to the proportion in a collection of Internet language.

$$Proportion = \frac{Count\ Modal\ Particle}{Count\ und\ and\ oder + Count\ Modal\ Particle} \tag{1}$$

An example illustrates the approach: *bloß* (mere) appears 387,392 times in the DECOW14AX corpus, while *und* and *oder* appeared in sum 308,628,935 times. Thus, a proportion of 0.0013 results. The idea is that when less common ground is present fewer modal particles will be used. The amount of *und* and *oder* in this case can therefore either increase or stay the same. In both cases the proportion will decrease. If people are referring to a common ground and thus use more modal particles, the frequencies of *und* and *oder* can either decrease or stay the same. In both cases the proportion will increase. We note the small resulting proportions, which result in very small standard errors and hence narrow confidence intervals.

2.2 Tabulation

The occurrences of the modal particles and the words *und* and *oder* in the different data sets were counted using R. The proportions were then calculated for each modal particle in each data set as described in Sect. 2.1. This led to 15 proportions per corpus, 60 respectively, to compare with the baseline proportions.

2.3 Statistical Analysis

Because we are dealing with big data, each modal particle proportion in a data set was compared to the baseline proportion using effect size metrics. Effect sizes and their surrounding 95% confidence interval were calculated in R using the Cox Logit method [20]. An effect size became significant if the 95%-CI excluded 0.

Furthermore, resulting effect sizes for the elections were compared to effect sizes of the comparison group. Deviations in the *d*-values were considered significant in cases where the lower bound of the 95%-CI for higher *d*-values and the 95%-CI upper bound for lower *d*-values did not overlap. Relative to significance testing, this approach is rather conservative.

3 Results and Discussion

As in Hampton and Shalin [1], the observed proportions of the modal particles in the event corpora as well as in the "nothing" corpora are highly correlated, despite adjustments for the influence of a common baseline (see Table 3). This result fails to distinguish language style during the election from language usage during "nothing".

Table 3. Partial spearman rank correlations between proportions controlling for normative influence.

	Elections 2	Elections 3	Nothing
Elections 1	.82*	.68*	.68*
Elections 2		.73*	.70*
Elections 3			.43

Note. $N = 15$ for all comparisons. *$p < .05$.

Effect size analyses are more informative. All calculated effect sizes deviate significantly from the baseline (see Table 4). Positive *d* indicate an increase in the modal whereas negative *d* indicate a decrease. Table 4 also shows the effect sizes in the election that differ significantly from the "nothing" corpus effect sizes. Moreover, effect sizes increase for the three election related events relative to the "nothing" data in 29 out of 45 cases ($P(K \geq 29, n = 45, p = 0.5) = 0.036$). Another 7 comparisons go in the same direction but were not significant, i.e., CIs overlapped. Just 6 modal particles decreased and are thus contradictory. Twenty four percent of the 45 effect sizes in the election corpora are small—below 0.20. Nearly 50% of the 15 effect sizes in the "nothing" corpus are below 0.20. This provides evidence for engagement specific to the election corpora.

Table 4. Effect size departures from norm by election event and control group.

Modal particle	Significant effect sizes *d*			
	Elections 1	Elections 2	Elections 3	Nothing
aber	0.20	0.22	**0.14**	0.19
auch	**0.04**	**0.10**	**0.02**	−0.02
bloß	**0.67**	**0.71**	**0.62**	0.35
denn	**0.13**	**0.11**	**0.14**	−0.19
doch	**0.40**	**0.51**	**0.45**	0.18
eben	**0.40**	**0.35**	**0.29**	0.14
eigentlich	**0.45**	0.44	**0.49**	0.39
etwa	−0.44	−0.44	−0.51	−0.50
halt	**0.61**	0.76	**0.53**	0.78
ja	**0.51**	**0.59**	**0.72**	0.43
mal	**0.43**	**0.43**	**0.37**	0.50
nur	**0.35**	**0.40**	**0.31**	0.21
schon	**0.34**	**0.39**	**0.34**	0.24
vielleicht	0.14	0.13	0.12	0.12
wohl	**0.48**	**0.57**	**0.48**	0.12

Note. All shown *d* are significant relative to norms. Positive *d* indicate an increase in the modal particle whereas negative *d* indicate a decrease. Bold *d* for the elections 1–3 are significantly different from the "nothing" *d* as indicated by non-overlapping confidence intervals.

4 Contributions

Twitter exchange, as measured by the presence of conversational modal particles, does differ from broad Internet language. The preponderance of significant differences in all corpora using conversational words is consistent with the view that Twitter constitutes a form of conversation [5]. However, modal particles in the three 2017 German election events also generally increased relative to the "nothing" corpus. Thus, these departures from Internet standards are not simply an artifact of Twitter. We interpret these indicators of common ground, point of view and anaphora as a measure of public engagement in a common event. Based on the observed differences, exchange regarding the German elections constitutes a collective narrative relative to an exchange with respect to "nothing".

Moreover, typically unexploited stop words are surely not nothing. In lieu of computational data driven methods, we employ linguistic, psycholinguistic and psychological theory to pre-select (and therefore interpret) our feature set. Our analysis of stop words expands the metrics of general social media analysis, providing a general feature that, now identified, could be combined with more conventional computational text mining. Stop words contain meaning; they need not be ignored. As in [5], focusing also on style and structure rather than only content can provide a first step of data analysis, of relevance to mining public opinion regarding virtually any consequential topic such as harassment, immigration, global warming or disaster response. Like sentiment, our metric is domain independent. Unlike sentiment analysis, our approach comes with an

underlying statistical and social science rationale that assists in interpretation, facilitating the comparison of engagement between events.

Acknowledgments. The first author recognizes support from the German Academic Exchange Service (DAAD). The third author's participation in this research was partially supported by NSF grants CNS 1513721 and EAR 1520870.

References

1. Hampton, A.J., Shalin, V.L.: Sentinels of breach: lexical choice as a metric of urgency. Hum. Factors **59**(4), 505–519 (2017). Davis, K. (ed.) Special Issue on Big Data/Winner of the 2016 Human Factors Prize
2. Thurmair, M.: Modalpartikeln und ihre Kombinationen. De Gruyter, Berlin (1989)
3. Bross, F.: German modal particles and the common ground. Helikon. Multidiscip. Online J. **2**, 182–209 (2012)
4. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124. Association for Computational Linguistics, Stroudsburg (2005). https://doi.org/10.3115/1219840.1219855
5. Purohit, H., Hampton, A., Shalin, V.L., Sheth, A.P., Flach, J.M., Bhatt, S.: What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination. Comput. Hum. Behav. **29**(6), 2438–2447 (2013)
6. Purohit, H., Hampton, A., Bhatt, S., Shalin, V.L., Sheth, A.P., Flach, J.M.: Identifying seekers and suppliers in social media communities to support crisis coordination. Comput. Support. Coop. Work (CSCW) **23**(4–6), 513–545 (2014)
7. Purohit, H., Banerjee, T., Hampton, A., Shalin, V.L., Bhandutia, N., Sheth, A.: Gender-based violence in 140 characters or fewer: a #BigData case study of Twitter. First Monday **21**(1–4) (2016)
8. Ebrahimi, M., Yazdavar, A.H., Sheth, A.: On the challenges of sentiment analysis for dynamic events. IEEE Intell. Syst. **32**(5), 70–75 (2017)
9. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. J. Am. Soc. Inform. Sci. Technol. **61**(12), 2544–2558 (2010)
10. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in Twitter events. J. Am. Soc. Inform. Sci. Technol. **62**(2), 406–418 (2011)
11. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends® Inf. Retr. **2**(1–2), 1–135 (2008)
12. Weydt, H.: Abtönungspartikel: die deutschen Modalwörter und ihre französischen Entsprechungen. Gehlen (1969)
13. Weydt, H., Hentschel, E.: Kleines Abtönungswörterbuch. In: Weydt, H. (ed.) Partikel und Interaktion, pp. 3–24. Niemeyer, Tübingen (1983)
14. Helbig, G.: Lexikon deutscher Partikeln. Verlag Enzyklopädie (1988)
15. Diewald, G.: Abtönungspartikel. In: Hoffmann, L. (ed.) Handbuch der deutschen Wortarten, pp. 117–142. De Gruyter, Berlin, New York (2007)
16. Götze, M., Geyer, S.: https://solariz.de/de/downloads/6/german-enhanced-stopwords.htm. Accessed 10 Mar 2018

17. Sheth, A., Jadhav, A., Kapanipathi, P., Lu, C., Purohit, H., Smith, G.A., Wang, W.: Twitris: a system for collective social intelligence. In: Alhajj, R., Rokne, J. (eds.) Encyclopedia of Social Network Analysis and Mining, pp. 2240–2253. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-6170-8_345
18. Schäfer, R.: Processing and querying large web corpora with the COW14 architecture. In: Bański, P., Biber, H., Breiteneder, E., Kupietz, M., Lüngen, H., Witt, A. (eds.) Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3). IDS, Lancaster (2015)
19. Schäfer, R., Bildhauer, F.: Building large corpora from the web using a new efficient tool Chain. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 486–493. European Language Resources Association (ELRA), Istanbul (2012)
20. Lipsey, M.W., Wilson, D.B.: Practical Meta-Analysis. Sage Publications Inc., Thousand Oaks (2001)