Triangulated Sentiment Analysis of Tweets for Social CRM

Simone E. Griesser
School of Applied Psychology
FHNW University of Applied Sciences and Arts Northwestern Switzerland
Olten, Switzerland
simone.griesser@fhnw.ch

Neha Gupta
Warwick Institute for the Science of Cities
University Of Warwick
Coventry, UK
Neha.Gupta@warwick.ac.uk

***Abstract*—High resolution data from social media platforms like *Twitter* presents an unprecedented opportunity to organisations for social customer relationship management (Social CRM) by analysing the ongoing discussion about business events such as a service outage. Text based sentiment analysis has been widely researched utilising mainly lexicon-based and machine learning approaches to uncover customers opinions. They are similar in the sense that the machine learning approach relies on an initial lexical model on which the learning is based. Both methods view sentiment as either positive, neutral, or negative. This is not the case for the psycholinguistic approach following which text sentiment is more continuous. We compare these three approaches with a Twitter dataset collected during a service outage. Contrary to our expectation, we find that the language used in tweets is not very negative or emotionally intense. This research therefore contributes to the sentiment analysis discussion by dissecting three methods and illustrating how and why they arrive at differing results. The selected research context provides an illuminating case about service failure and recovery.**

***Index Terms*—Text Analysis, Sentiment Analysis, Psycholinguistics, Valence, Arousal, Emotional Intensity, Social CRM, Twitter, NLP, Service Breakdown**

I. Introduction

Social media has brought many innovations, one of which is the measurement of customer sentiment. The analysis of customer language for sentiment has been largely driven by computer scientists and gained widespread implementation in industry. Sentiment in a nutshell refers to positive or negative valence. The digitally connected society enables customers to openly and instantly share their opinions on social media. This has paved the way for social CRM into main stream CRM and has made it pivotal for any business strategy [1] to monitor customer opinions and proactively manage them. The monitoring of customer views for example, is positively related to customer relationship performance [2]. Hence successful customer relationships rely on understanding customer sentiment in order to be able to address them effectively. Sentiment analysis incorporates lexical and machine learning as an underlying mechanisms to compute a sentiment score of customer opinions on social media sites [3]. Psychologists have studied sentiment, which they call valence or mood, for much longer. In fact, a benchmark comparison by Ribeiro et. al. [4] of 24 popular sentiment analysis methods shows how methods that were originally developed in linguistics and psychology are increasingly used in computational sentiment analysis. Using the rules of language, the field of psycholinguistic examines language comprehension and investigates the relationship between language and psychological processes. In comparison to psycholinguistics, lexical and machine learning approaches tend to see sentiment as a quasi-dichotomy. In other words, sentiment is either positive, negative, or neutral. Considering how we as individuals experience emotions, we intuitively know that any emotional experience is not as clear cut. We experience any positive or negative emotion on a continuum. Emotions not only vary in terms of their valence, i.e. positive or negative, but also in terms of their intensity. For example, we might feel content, happy, or joyous. These are all positive emotions, but differ in intensity. Therefore, any emotional experience is two-dimensional [5].

In this paper we compare traditional computer science sentiment analysis techniques with the psycholinguistics method to better understand the similarities and differences of these approaches, which ultimately influence the inferences that we can draw. In fact, previous researchers highlighted that the methodologies in these domains largely complement each other [6]. We use a Twitter dataset collected during the Skype outage, which provides a novel context. To the best of our knowledge, computer science and psycholinguistics approaches to sentiment analysis have not been empirically compared. Similarly, sentiment has not been studied during a service outage. This paper therefore not only fills a methodological gap, but also speaks to industry in two important ways. First, a solid understanding about sentiment methods is required to derive reliable insights, which are key for sound decision making. Second, service recovery is crucial for customer satisfaction, loyalty, and trust.

II. Data and Methodology

A. Twitter data collection

The dataset used in this paper was collected using Twitter streaming API on 21 September 2015 after observing the Skype Twitter account notifying customers about their messenger service outage [7]. The tweets were searched and collected using two keywords, '#skypedown' and 'skypedown' in the tweet text. Approximately 10,000 tweets were collected for the day using twitter4j API Java package.

978-1-7281-3105-4/19/$31.00 ©2019 IEEE
DOI 10.1109/SDS.2019.000-4

B. Tweets sentiment scoring

We scored the tweets text using three natural language processing (NLP) approaches - lexical, machine learning, and psycholinguistic. The advantage of using lexicon based methods for sentiment analysis is that it does not require training data and is often claimed to be successful for domain independent sentiment classification. The *lexical* scoring algorithm first pre-processed the tweet by incorporating the tokenization process (removing punctuation, converting to lower case, and removing stop words etc.). We then extracted the unigram (single word) features to understand the polarity of each tweet word. To do so, we matched each unigram with the Bing-Liu opinion lexicon [8], a corpus of positive and negative words, and arrived at the tweet sentiment score by subtracting the number of negative word occurrences from the number of positive word occurrences for each tweet. With this we arrived at the first simple measure of tweet sentiment. In order to utilise *machine learning* approach, we resorted two labelled datasets for training and testing, provided by SemEval2015 (Semantic Evaluation), which is an ongoing series of NLP competitions [9], to arrive at a machine learning score for Skype outage tweets. Both datasets have tweets annotated with three sentiment categories: positive, negative and neutral. Using the training dataset from SemEval2015, we classified our unlabelled tweets about the Skype outage. In this classification we used the features part of speech tags, word vectors, unigrams, bigrams, and sentiment lexica. Sentiment lexica provided us with the objectivity or subjectivity of matched words present in the lexica and we provisioned four lexicas for this study; Bing Liu opinion lexicon [8], the MPQA subjectivity lexicon [10], AFINN [11] and SentiWordNet [12]. We then used a logistic regression classifier to arrive at machine learning sentiment score (positive, negative and neutral) by incorporating the methodology adopted in our previous research [13].

For the psycholinguistics analysis, the tweets were read into the R environment where they underwent a data cleaning process. We removed numbers, website links, emoticons, special characters, and stop words from tweets with the packages tm and NLP from the CRAN repository in R. The remaining linguistic content of tweets was rated according to word valence and arousal ratings found in a published database containing 13,915 word lemmas [5]. This database has, for example, been used in the marketing and behaviour disciplines [14] [15]. The norms have been collected in a crowd-sourcing effort. Each word was rated by 18 different participants on a nine-point scale ranging from one (unhappy; calm) to nine (happy; excited). Participants indicated how they felt when reading a word. One end of the valence scale was anchored with completely unhappy, annoyed, unsatisfied, melancholic, or despaired. The other end of the scale was labelled completely happy, pleased, satisfied, or contented. (see [13] page 5 for details). This detailed scale anchoring enables a more complete valence rating that clearly defines positive sentiment as opposed to scales that use only happiness as anchoring points. In a final step, the mean and median sentiment per tweet were computed based on the valenced rated words in order to have two complimentary measures of dispersion because natural language data is not always normally distributed.

C. Comparing and Contrasting the Lexical, Machine Learning, and Psycholinguistic Approaches

Comparing and contrasting the lexical, machine learning, and psycholinguistic approaches provides different predictions. The lexical and psycholinguistic approaches are insofar similar as they both rely on unigrams and use a single lexicon. In comparison, machine learning uses bigrams and draws on more lexica. We might thus expect the sentiment scores of the lexical and psycholinguistic approaches to be similar. However, they use different lexicons to look up sentiment scores. The employed lexical and machine learning approaches on the other hand, rely on the same initial lexicon. Following this rationale we could expect the sentiment scores of the lexical and machine learning approaches to agree with each other. Similarly, the lexical and machine learning approaches establish sentiment by weighing positive and negative words against each other because the number of negative word occurrences are subtracted from the number of positive word occurrences. The word ‘delighted’ is therefore treated as equally positive as the word ‘content’. Similarly, the word ‘upset’ is treated as equally negative as the word ‘displeased’. In sum, the three chosen approaches share characteristics but also deviate from each other making them an interesting and suitable choice for exploratory comparisons.

III. Results

The initial data exploration shows that sentiment depends on the employed approach as Fig. 1 shows. The lexical approach shows that generally neutral language is employed. Following the machine learning approach, tweeters use almost as much neutral as negative language. According to the psycholinguistic approach, tweeters use on average slightly positive language. The same holds true for the median sentiment per tweet. Fig. 1 illustrates the quasi-dichotomy of the lexical and machine learning approaches and the more detailed analysis that the psycholinguistic approach allows. It therefore underpins our motivation to study the different approaches. Next, we compute Kendall’s *tau* as a correlation measure.

According to Table. 2, the lexical and machine learning approaches share a moderate positive correlation. The lexical and psycholinguistic approach share an equally moderate correlation, i.e. produce similar results. In comparison, the results obtained with the machine learning and the psycholinguistic approaches share less agreement. The mean and median psycholinguistic methods strongly correlate with each other emphasising the robustness of the psycholinguistic approach. These similarities or differences cannot be explained in terms of data cleaning processes or differing stop words. All three approaches rely on the same stop word list. The reasonable correlation between the lexical and psycholinguistic approach

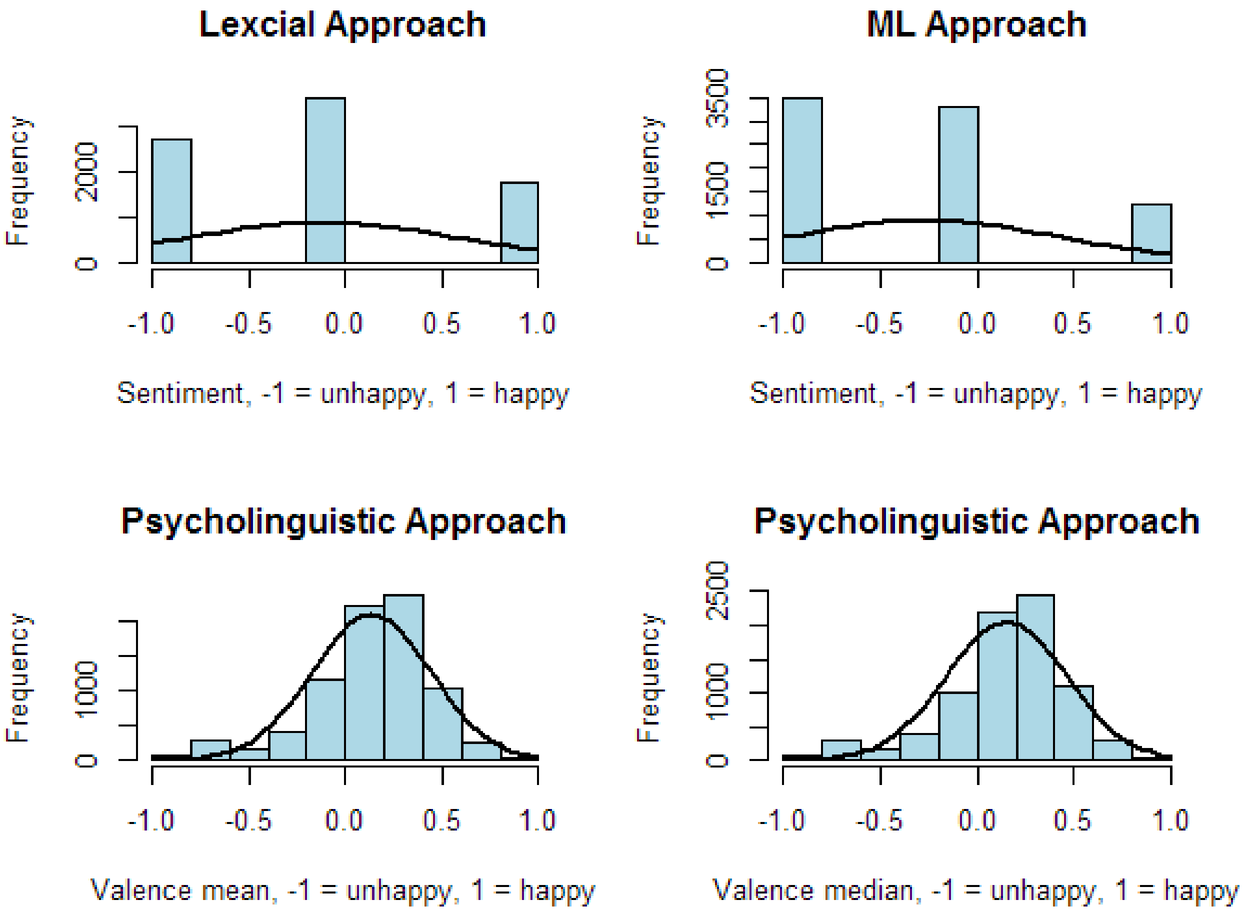
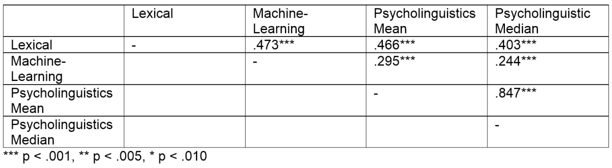


Fig. 1. Triangulated sentiment analysis for Skype outage tweets



	Lexical	Machine-Learning	Psycholinguistics Mean	Psycholinguistic Median
Lexical	-	.473***	.466***	.403***
Machine-Learning		-	.295***	.244***
Psycholinguistics Mean			-	.847***
Psycholinguistics Median				-

*** p < .001, ** p < .005, * p < .010

Fig. 2. Correlation Analyses with Kendalls *tau*

suggests that employed dictionary databases are similar. The much smaller correlation between the psycholinguistics and machine learning approaches therefore indicates that the approaches start to deviate from each other with the learning algorithm. The moderate correlation between the lexical and machine learning approaches support this notion.

A. The Nuances of Psycholinguistics

Particularly in the context of service outage, customer sentiment is a useful indication how effective the service recovery efforts are. The intensity with which customers experience positive or negative sentiment however is equally important. Based on the customer delight notion [16], delighted customers are more satisfied and loyal than content customers. Delight expresses positive emotion more strongly than content. Hence, strongly experienced emotions more powerfully influence customer satisfaction than weakly experienced emotions. The lexical and machine learning approaches treat the words ‘delighted’ and ‘content’ as equally positive. According to psycholinguistics, the word ‘delighted’ has a much higher positive valence rating, i.e. 7.82, than the word ‘content’, i.e. 6.70. In that sense, the psycholinguistics valence ratings answer to this issue, because it computes the average or median sentiment per text unit. In addition, psycholinguistics provides information on how emotionally intense the language

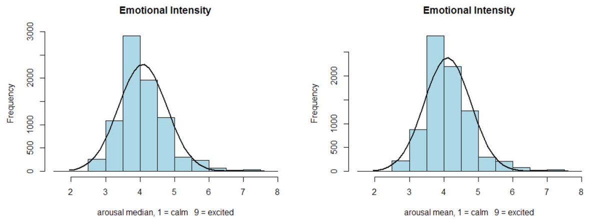


Fig. 3. Language arousal as a proxy for emotional intensity

is with language arousal ratings. Language arousal is part of the same dictionary as language valence [5]. The language arousal scale thus ranges from one (calm) to nine (excited) and the vast majority of words were rated by at least 18 different individuals.

We explore the emotional intensity notion by computing the mean and median language arousal for each tweet. As Fig. 3 shows, customers used only slightly arousing language when tweeting about the Skype outage. This is surprising given that tweets concern a service failure. A possible explanation is that only very unhappy customers expressed the emotion more intensely. In fact, very unhappy customers with a strong emotional experience are a segment that companies need to identify, monitor, and manage carefully. In order to tackle this concern, very unhappy and very happy tweets were identified by selecting the tweets in the sample that were three standard deviations above and below the mean. The sentiment of these tweets is then correlated with language arousal. With increasing positive sentiment, unhappy customers use slightly calmer language, but the correlation between sentiment and arousal is weak (*tau* = - .115, z = -15.453, p <.001). The analysis of happy customers paints a similar picture. The correlation between positive sentiment and arousal is weak (*tau* = - .143, z = -5.185, p <.001). Therefore, very unhappy and very happy customers do not express the emotions triggered by the Skype service outage considerably more or less strongly.

IV. Limitations and further ongoing work

This work is the first of its kind and exploratory in nature. Many different lexica, i.e. sentiment dictionaries, exist (see for example [5] [8] [17]). Many computer science lexica do not provide information about how many individuals rated each word in the lexica, what the rating instructions were, and who the rating individuals are. Moreover, most lexica can be adapted by the researcher to fit different research needs. The differences between lexica and how they impact accuracy are unclear. Similarly, how geographical consistency in training data and differences in learning algorithms impact sentiment score is underexplored. These are important questions because they impact the reliability of sentiment score. Overestimating or underestimating positive sentiment in customer language about a trend, for example, potentially results in misallocating resources in the short term and impacting market share in the long term. Hence, reliable information are the basis of good business decisions.

V. Conclusions

In this initial investigation, we compared the lexical, machine learning, and psycholinguistic approaches to ascertain how Skype customers felt during the Skype outage. Lexical and machine learning, the traditional approaches in computer science, are moderately correlated at best. The learning algorithm and the training data thus account for almost 60% of the deviation between the lexical and machine-learning approaches. A moderate correlation exists between the lexical and psycholinguistic approach. The machine-learning and psycholinguistic approach are weakly correlated and produce very different results. These findings raise interesting questions with reference to the incremental value of learning algorithms and training data. Limited knowledge and awareness exists about these issues despite their importance. Sentiment computation tends to be a black box for many industry users who take important business decisions. The sentiment computation procedure influences not only the results but also the inferences we can draw from the. For example, the extent to which the training data takes regional differences into account is unclear. English is the first language of the United Kingdom, the United States of America, and Australia. However, British, American, and Australian English each have their own expressions. More importantly, English is spoken across the globe. Non-native speakers may use the language differently. More careful consideration must therefore be given to language regions when using lexical databases and training data to ensure consistency and ultimately comparability.

Contrary to our expectation, the findings reveal that the language in tweets mentioning the Skype outage was not very negative. Instead the language was slightly positive and only a little emotionally intense suggesting that customers were not strongly upset about the outage. There are different possible explanations for this. Skype was either very good at service recovery, customers have lower expectations towards a service that is free, or customers employed coping strategies effectively. More research is required to better understand what made the Skype service recovery successful, i.e. why it did not trigger more negative and strong responses from customers. In conclusion, psycholinguistics provides a more transparent method for sentiment analysis and affords additional insight about emotional intensity that the traditional sentiment analysis approaches cannot provide.

Acknowledgements

We thank David Purser, Warwick Institute of Science of Cities for helping us to arrive at the machine learning sentiment score for tweets. We also thank Prof. Stephen Jarvis, Department of Computer Science, University of Warwick to provide resources to conduct this study. This research is partially supported by the EPSRC Centre for Doctoral Training in Urban Science and Progress (EP/L016400/1)

References

[1] N. Woodcock, A. Green, and M. Starkey, “Social crm as a business strategy,” *Journal of Database Marketing & Customer Strategy Management*, vol. 18, no. 1, Mar 2011. [Online]. Available: https://doi.org/10.1057/dbm.2011.7

[2] K. J. TrainoraJames, J. M. Andzulis, A. Rapp, and R. Agnihotri, “Social media technology usage and customer relationship performance: A capabilities-based examination of social crm,” *Journal of Business Research*, vol. 67, no. 6, June 2014. [Online]. Available: https://doi.org/10.1016/j.jbusres.2013.05.002

[3] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste, “Twitter sentiment analysis: Lexicon method, machine learning method and their combination,” 2015. [Online]. Available: https://arxiv.org/pdf/1507.00955.pdf

[4] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto, “Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods,” *EPJ Data Science*, vol. 5, Jul 2016. [Online]. Available: https://doi.org/10.1140/epjds/s13688-016-0085-1

[5] A. B. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 english lemmas,” *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, Dec 2013. [Online]. Available: https://doi.org/10.3758/s13428-012-0314-x

[6] M. Brysbaert, E. Keuleers, and P. Mandera, “A plea for more interactions between psycholinguistics and natural language processing research,” *Computational Linguistics in the Netherlands Journal*, vol. 4, p. 14, 2014.

[7] T. Telegraph, “Skype outage sees internet calls go down in many countries,” 2015. [Online]. Available: https://www.telegraph.co.uk/technology/news/11879664/Skype-outage-sees-internet-calls-go-down-in-many-countries.html

[8] Bing-Liu. (2004) Opinion lexicon. [Online]. Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

[9] S. R. et. al. (2015) Semeval 2015 task 10. [Online]. Available: http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools

[10] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 347–354.

[11] F. Nielsen. (2011) Afinn. Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?6010

[12] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.” in *LREC*, vol. 10, 2010, pp. 2200–2204.

[13] N. Gupta, H. Crosby, D. Purser, S. Jarvis, and W. Guo, “Twitter usage across industry: a spatiotemporal analysis,” in *IEEE International Conference on Big Data Computing Service and Applications*, March 2018.

[14] J. Ren and J. V. Nickerson, “Online review systems: How emotional language drives sales,” April 2014. [Online]. Available: https://ssrn.com/abstract=2426694

[15] D. Hildebrand, Y. DeMotta, S. Sen, and A. Valenzuela, “Consumer responses to corporate social responsibility (csr) contribution type,” *Journal of Consumer Research*, vol. 44, no. 4, pp. 738–758, 2017. [Online]. Available: http://dx.doi.org/10.1093/jcr/ucx063

[16] R. L. Oliver, R. T. Rust, and S. Varki, “Customer delight: foundations, findings, and managerial insight,” *Journal of Retailing*, vol. 73, pp. 311–336, 1997.

[17] M. Thelwall, “Heart and soul: Sentiment strength detection in the social web with sentistrength,” *Proceedings of the CyberEmotions*, pp. 1–14, 2013.