A Comparison of Lexicon-Based and ML-Based Sentiment Analysis: Are There Outlier Words?

Siddhant Jaydeep Mahajani[1], Shashank Srivastava[1] and Alan F. Smeaton[1,2]
[1]School of Computing and [2]Insight Centre for Data Analytics
Dublin City University, Glasnevin, Dublin 9, Ireland
email: alan.smeaton@dcu.ie

***Abstract*—Lexicon-based approaches to sentiment analysis of text are based on each word or lexical entry having a pre-defined weight indicating its sentiment polarity. These are usually manually assigned but the accuracy of these when compared against machine leaning based approaches to computing sentiment, are not known. It may be that there are lexical entries whose sentiment values cause a lexicon-based approach to give results which are very different to a machine learning approach. In this paper we compute sentiment for more than 150,000 English language texts drawn from 4 domains using the Hedonometer, a lexicon-based technique and Azure, a contemporary machine-learning based approach which is part of the Azure Cognitive Services family of APIs which is easy to use. We model differences in sentiment scores between approaches for documents in each domain using a regression and analyse the independent variables (Hedonometer lexical entries) as indicators of each word's importance and contribution to the score differences. Our findings are that the importance of a word depends on the domain and there are no standout lexical entries which systematically cause differences in sentiment scores.**

I. Introduction

Sentiment analysis is an established form of text analysis which measures to what extent the sentiment behind a piece of text is positive, negative or neutral. The most popular implementation combines machine learning and natural language processing (NLP) though one of the downsides is domain-dependence where classifiers need to be attuned to different text domains. An alternative approach is lexicon-based using a dictionary of words with pre-defined sentiment ratings. This has the advantage of domain-independence but brings a disadvantage and a commonsense assumption that the semantics of a word should depend on itself and also its use and context. The use of word context can give significant improvements on a wide range of NLP tasks including sentiment analysis.

Here we are interested in how domain-independent are lexicon-based sentiment analysis tools, how do the "baked-in" word-level sentiment weights contribute to differences when compared to machine learning approaches, and how transferable are they across domains? We take a popular lexicon-based sentiment analysis tool, the Hedonometer [1], and compare its output against that from a popular machine learning based tool, Microsoft Azure's Text Analytics technology [2] on collections of text from four domains. We set the Azure sentiment analysis as a standard to aim at and we use a regression to model the differences in sentiment analysis from the two approaches across each of the four domains. We then examine significance values for the variables from the regression thus revealing what are the lexical entries, i.e., words which have the greatest and least impact on differentiating between Hedonometer and Azure sentiment scores. This highlights Hedonometer words whose sentiment weights may need to be updated and those which should be left untouched if we wanted Hedonometer sentiment to match Azure sentiment, for each domain. The insights this provides will help us understand the strengths and limitations of lexicon-based approaches to calculating sentiment scores.

II. Background and Related Work

A. Sentiment Analysis

Sentiment analysis, or opinion mining, is the automatic analysis of text in order to determine the attitude or judgement that the text prompts in a typical reader [3]. It has widespread use in social media monitoring, brand monitoring and reputation management, product analysis and customer reviews, and in market research [4]. The most popular implementation combines machine learning and one of its sub-fields, NLP, on manually annotated training data to achieve systems which are robust and scalable [5]. One of the downsides of such approaches is domain-dependence where "linguistic and content peculiarities require a domain-specific sentiment source" [6].

B. Lexicon-Based Sentiment Analysis

An alternative approach to using machine learning is word or lexicon-based where the polarity of a text is determined by searching for words or phrases which have predetermined weights as indicators of sentiment, then combining the word weights in some way. Lingmotif is a lexicon-based, linguistically-motivated, sentiment analysis tool [7] which performs analysis on input text based on the identification of sentiment-laden words and phrases from Lingmotif's rich core lexicons. It also employs context rules to account for sentiment shifters. SentText is another tool for lexicon-based sentiment analysis [8] which performs sentiment analysis with predefined sentiment lexicons or self-adjusted lexicons. Finally, Syuzhet is a lexicon-based tool for sentiment analysis of literary texts that draws upon the Syuzhet, Bing, Afinn, and NRC lexicons [9] containing 10,748, 6,789, 2,477 and 13,901 words respectively.

Despite shortcomings, lexicon-based sentiment methods are widely used. The methods have often been criticised for their accuracy but recent work [10] has shown that lexicon-based methods can work well where neither qualitative analysis such as manually assigned ground truth labels, nor a machine learning-based approach, is possible.

C. Hedonometer

The Hedonometer is a lexicon-based sentiment analysis tool which measures average “happiness” or sentiment using a lexicon of 10,222 common words in the English language, each of which has a context-free estimation of its “happiness” score. These scores were calculated using a language assessment Mechanical Turk where users were asked to rate each word on a nine-point integer scale according to how it made them feel [1]. Examples of the averaged scores for some words from the Hedonometer are shown in Table I.

TABLE I
Sample from the 10,222 words and their scores in the range 1 (sad) to 9 (happy) from [11]

Word	Score	Word	Score	Word	Score
laughter	8.50	food	7.44	reunion	6.96
the	4.98	of	4.94	vanity	4.30
hate	2.34	funeral	2.10	terrorist	1.30

Since its introduction, the Hedonometer has demonstrated stability and reliability along with a remarkable quality of tunability [11]. The algorithm to compute sentiment for a document initially extracts the frequencies of occurrence and then the average sentiment from a given text which is subsequently normalised for document length.

Many lexicon-based sentiment analysis tools such as Hedonometer have limitations as they fail to consider word context. For instance, the phrase *not happy* would receive a positive sentiment score, but the phrase *not unhappy* would receive a negative one. Hence we can say that Hedonometer, like most lexicon-based approaches to sentiment analysis, should not be very reliable in calculating sentiment scores for short texts but if errors in sentiment are not connected then this may not be an issue when it is used on a large quantity of text [12].

III. Data Collections

For the analysis of Hedonometer vs. Azure sentiment analysis we used annotated English language data sets from four domains: Finance, News, Social Media, and IMDb customer reviews. The premise is that data from different domains helps us to identify the most commonly used words in that domain, i.e., domain-specific set of words as well as the words that are common irrespective of what their domain is.

Finance: Finance is a domain where sentiment is important as it can influence stock market trends. We used data from [13] who used it to identify semantic orientations in economic texts. It consisted of c.5,000 phrases/sentences sampled from financial news texts and company press releases, tagged as positive, negative or neutral.

News: The news data set was used by [14] to perform sentiment analysis on news that are displayed everywhere. It consists of almost 50,000 news articles for an 8 month period from November 2015 to July 2016 on four different topics: economy, Microsoft, Obama and Palestine.

Social Media: The social media data was a collection of 40,000 tweets and used by [15] to perform sentiment analysis on Tweets posted by users with the specific task of emoji prediction.

IMDb Reviews: The IMDb customer reviews consisted of 50,000 reviews posted on IMDb, an International Movie Database platform. The data set was used by [16] to perform sentiment analysis.

The number of Hedonometer lexical entries appearing across texts in each domain is shown in Table II showing that texts from the news domain have much fewer Hedonometer words. However the reader is reminded that our objective is to see if it is possible to identify lexical entries in the Hedonometer which cause it to differ from a machine learning based approach and not to determine the ultimate adjustments to Hedonometer weights that should be enacted.

TABLE II
Charactistics of Hedonometer words from the lexicon of 10,222, appearing in datasets from each domain

	Finance	News	Soc. media	IMDB	All domains
# Hedonometer words	966	274	1,886	2,673	3,810

Numbers of words appearing in any of			
1 domain	2 domains	3 domains	all domains
2,396	941	371	102

Calculating Sentiment Scores: we used Microsoft Azure’s Text Analytics technology [2], an established machine-learning based sentiment analysis tool, against which to match Hedonometer scores. This scores texts in the interval 0 (negative) to 1 (positive) for sentiment. We computed the sentiment score for each document in each domain using Hedonometer and Azure and Figure 1 shows, for each domain, the differences in scores. For finance and news there is a relatively flat part of the graph in the middle indicating approximate agreement between Hedonometer and Azure, with small numbers of documents at each end where there are larger differences in sentiment. For social media documents there are greater differences between Hedonometer and Azure ratings while for the IMDB reviews there are extreme differences with few documents having agreed or even close scores (the crossover part of the graph). This is caused by a relatively small range of values from the Hedonometer (shown in pink) while the Azure ratings have a much greater range of values.

IV. Results and Analysis

We set the Azure sentiment scores as a target and use linear regression to model the differences in results between the Hedonometer and Microsoft Azure’s Text Analytics technology across texts from each of the four domains. We then examine significance values for variables from the regression

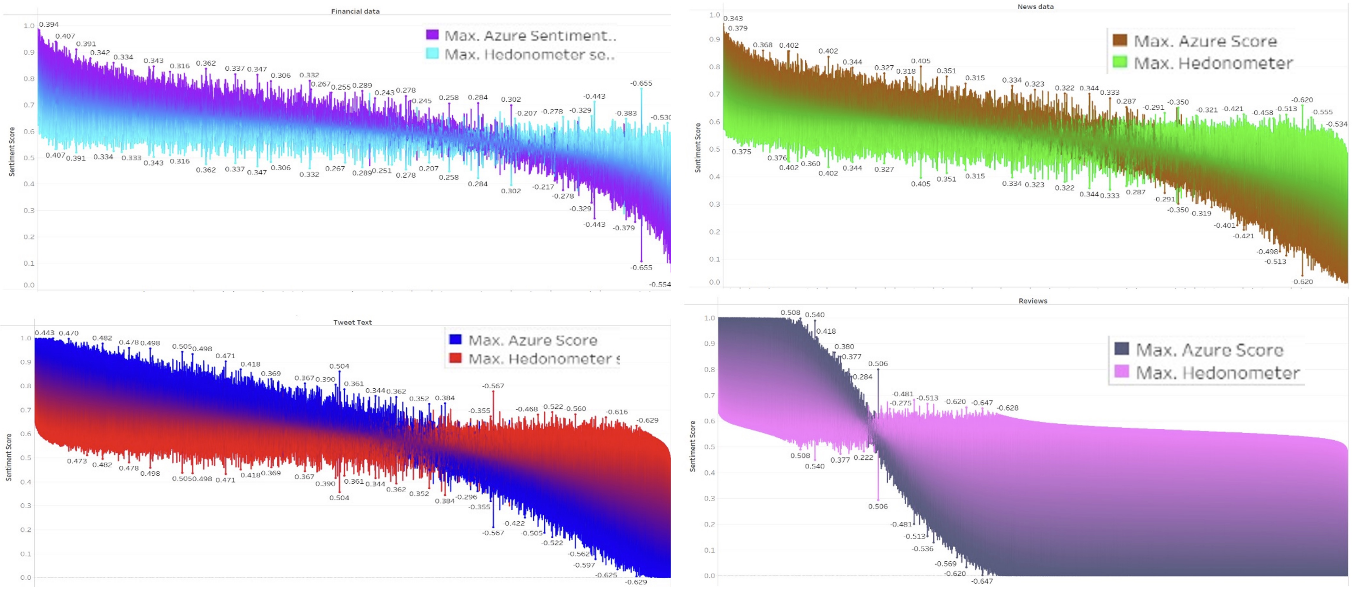

Fig. 1. Differences between Hedonometer and Azure sentiment scores for finance, news, social media and IMDB review domains, respectively.

revealing which words have greatest and which have least impact on differentiating between Hedonometer and Azure. The top half of Table III lists words from each domain and from a combination across all domains, with the smallest p-values indicating words whose contributions are different between Hedonometer and Azure. These are words whose sentiment weights would need to be changed to make the Hedonometer more like Azure. The bottom half of Table III lists words with the largest p-values indicating words with the same interpretations in Hedonometer as in Azure. We limit our analysis to words which appear in Hedonometer's lexicon and in at least three of the four text domains to see if there are words consistently outliers across domains.

We measured the correlation between rankings of (some) Hedonometer words by their "happiness" scores vs. the p-values from differences between Hedonometer and Azure. Table IV shows those correlations with words whose p-values equal to 0 removed. This shows a negative correlation for Finance and News and a positive but not strong correlation for the others . What this means is that the importance of a word towards the differences between the two sentiment analysis approaches has only a small correlation with the Hedonometer ranking of that word.

V. Discussion and Conclusions

Though the number of Hedonometer words in each of the four domains may limit our analysis, restricting it to entries from just a portion of the possible lexicon, this does not detract from the process of trying to identify lexical entries which cause it's output to differ from the Azure machine learning approach. Table III indicates that when mapping Hedonometer sentiment to Azure sentiment, different words are more, and less, important for different domains. This may be because the sets of Hedonometer words appearing in the texts from the four domains are a fraction of the overall Hedonometer lexicon, 3,810 from 10,222 possible entries. Even so, there are no major outlier words that stand out which is surprising, but informative and the only word with smallest or largest p-value that appears in more than one domain is "mom". Our Spearman correlation between Hedonometer ranking and p-values for modelling the differences, bears this out. The nature of the words in Table III do not appear to be particularly domain-dependent and from those words it would be difficult to match them to their domain.

Our future work may include targeting the impact of specific Hedonometer words for their impact on sentiment analysis by using texts containing those words rather than using words from particular domains as we have done here. This would give us greater coverage than the 3,810 words we analysed here though the results may be the same. There may also be an issue around our use of linear regression to determine p-values in the situation where the many independent variables in the model, the lexical entries, lead to multiple hypothesis testing. While correction methods for multiple hypothesis testing exist [17] which we could use in future, an alternative would be to substitute the regression model with a non-parametric machine learning model where feature importance scores could be used to evaluate the importance of individual lexical entries.

Considering that there is no such thing as universally agreed sentiment scores [18] and even inter-annotator agreement among humans is only about 80%, adjustments to the weights of Hedonometer lexical entries would seem to make little or no difference to overall sentiment scores. In the actual use of sentiment analysis tools, so long as they are used consistently and any comparisons are like-with-like and not across sentiment analysis tools or approaches then modifications to weights in lexicon-based approaches may not be worthwhile.

TABLE III

HEDONOMETER LEXICON WORDS WITH SMALLEST AND LARGEST P-VALUES WHEN MODELLING HEDONOMETER VS. AZURE OUTPUTS.

	Individual domains				Combined domains
	Finance	News	Social media	IMDB reviews	ranked by p-value
Smallest p-values	jason	great	great	listen	understand
	sign	mom	wish	video	walk
	strong	water	may	dream	bad
	point	mail	feel	rigid	meant
	profit	video	wait	hang	end
	fan	press	want	ad	goodnight
	tough	strength	still	known	space
	owl	bomb	mom	avoid	main
	matt	earn	kind	violent	faith
	rate	flag	found	laura	sister
Largest p-values	mon	taken	louis	new	high
	greg	small	water	use	sport
	notion	consider	town	editor	sign
	co	india	ipad	charter	blank
	blank	reach	bout	paid	upset
	pilot	chose	demon	written	worst
	shall	worst	Monday	snake	new
	take	jordan	top	role	good
	bear	ok	round	ten	best
	report	upset	friend	sweden	opinion

TABLE IV

CORRELATION BETWEEN HEDONOMETER SCORE RANKING AND RANKING BY P-VALUE

Domain	Number of Hedonometer words used	Spearman correlation
Finance	808	-0.1126
News	259	-0.1206
Social media	967	0.2949
IMDB reviews	1,369	0.3300

Acknowledgment This work was part-funded by Science Foundation Ireland under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund. Data is available at https://doi.org/10.6084/m9.figshare.24539410.

REFERENCES

[1] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth, "Happiness and the Patterns of Life: A Study of Geolocated Tweets," *Scientific Reports*, vol. 3, no. 1, p. 2625, Sep. 2013. [Online]. Available: https://doi.org/10.1038/srep02625

[2] A. Carvalho and L. Harris, "Off-the-shelf technologies for sentiment analysis of social media data: Two empirical studies," in *Proceedings of the Twenty-Sixth Americas Conference on Information Systems (AMCIS)*, Salt Lake City, Utah, USA, 2020.

[3] M. Neethu and R. Rajasree, "Sentiment Analysis in Twitter using Machine Learning Techniques," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. IEEE, 2013, pp. 1–5.

[4] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[5] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, and A. Rehman, "Sentiment analysis using deep learning techniques: a review," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017.

[6] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial Intelligence in Medicine*, vol. 64, no. 1, pp. 17–27, 2015.

[7] A. Moreno-Ortiz, "Lingmotif: Sentiment analysis for the digital humanities," in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Apr. 2017, pp. 73–76.

[8] T. Schmidt, J. Dangel, and C. Wolff, "Senttext: A tool for lexicon-based sentiment analysis in digital humanities," in *Proc. 16th International Symposium of Information Science (ISI)*. Werner Hülsbusch, 2021.

[9] H. Kim, "Sentiment analysis: Limits and progress of the Syuzhet package and its lexicons." *Digital Humanities Quarterly*, vol. 16, 2022.

[10] E. Öhman, "The validity of lexicon-based sentiment analysis in interdisciplinary research," in *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. NIT Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 7–12. [Online]. Available: https://aclanthology.org/2021.nlp4dh-1.2

[11] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter," *PLoS ONE*, vol. 6, no. 12, p. e26752, 2011.

[12] J. Gibbons, R. Malouf, B. Spitzberg, L. Martinez, B. Appleyard, C. Thompson, A. Nara, and M.-H. Tsou, "Twitter-based Measures of Neighborhood Sentiment as Predictors of Residential Population Health," *PLoS ONE*, vol. 14, no. 7, p. e0219550, 2019.

[13] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts," *J. Assoc. for Info. Science and Technology*, vol. 65, 2014.

[14] D. Dangi, S. T. Chandel, D. K. Dixit, S. Sharma, and A. Bhagat, "An Efficient Model for Sentiment Analysis using Artificial Rabbits Optimized Vector Functional Link Network," *Expert Systems with Applications*, vol. 225, p. 119849, 2023.

[15] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Proceedings of Findings of EMNLP*, 2020.

[16] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011. [Online]. Available: https://aclanthology.org/P11-1015

[17] O. Menyhart, B. Weltz, and B. Győrffy, "Multipletesting.com: A tool for life science researchers for multiple hypothesis testing correction," *PLoS One*, vol. 16, no. 6, p. e0245824, 2021.

[18] C. S. Khoo and S. B. Johnkhan, "Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons," *Journal of Information Science*, vol. 44, no. 4, pp. 491–511, 2018. [Online]. Available: https://doi.org/10.1177/0165551517703514