The Dark Side of Sentiment Analysis: An Exploratory Review Using Lexicons, Dictionaries, and a Statistical Monkey and Chimp.

PREPRINT - Can be cited as: Samuel, J., Rozzi, R. and Palle, R. (2022), "The Dark Side of Sentiment Analysis: An Exploratory Analysis Using Lexicons, Dictionaries, and a Statistical Monkey and Chimp".

Jim Samuel
jim@aiknowledgecenter.com
Rutgers University

Gavin Rozzi
gavin.rozzi@rutgers.edu
Rutgers University

Ratnakar Palle
rrpalle@gmail.com
Apple Inc.

Abstract

This article discusses the inconsistencies, inaccuracies and challenges, namely the 'dark side' of sentiment analysis and then demonstrates problems with using sentiment analysis lexicons or dictionaries for estimating sentiment in textual artifacts. Sentiment analysis, an important dimension of natural language processing (NLP), has seen an exponential adoption rate across research and practitioner disciplines. Many interesting developments in NLP methods continue to improve the accuracy of sentiment analysis. However, the plethora of sentiment analysis methods, dictionaries and lexicons, tools, open source code for machine learning based sentiment analysis, and of-the-shelf sentiment analysis solutions have led to a flurry of research and applied solutions without sufficient concern for the limitations, context, and the inaccuracies of sentiment analysis, and the inherent ambiguities associated with the unaddressed sentiment analysis domain challenges. Scant attention is given, especially in applied research and industry usage, to the inherent ambiguities associated with the unanswered questions pertaining to the science of sentiment analysis. This study reviews known issues with sentiment analysis as documented by prior research and then compares the application of multiple of-the-shelf lexicon and dictionary methods to stock market and vaccine tweets. The intention is not in any way to improve the accuracy of sentiment analysis as compared to prior benchmarks but to identify and discuss critical aspects of the dark side and develop a conceptual discussion of the characteristics of the dark side of sentiment analysis. We conclude with notes on conceptual solutions for the dark side of sentiment analysis and point to future strategies that could be used to improve the accuracy of sentiment analysis and understanding. This research will also help align researcher and practitioner expectations to understanding the limits and boundaries of natural language processing based solutions for sentiment analysis and estimation.

1 Introduction

Sentiment analysis has become a prominent multidisciplinary research paradigm, as evidenced by its usage surge in scholarly articles: Google Scholar alone documents over 177,000 articles with the exact phrase "sentiment analysis", and over 2.5 million articles related to sentiment analysis! (Scholar, 2021). Sentiment analysis has been used for mining insights from a wide range of domains and topics such as politics (Britzolakis et al., 2020), responses to articles(Muddiman and Stroud, 2017), investor behavior in stock markets (Pelaez et al., 2021; Samuel, 2017a), COVID-19 reopening 'Public Sentiment Scenarios' (Samuel et al., 2020b), socioeconomic implications of COVID-19 (Rahman et al., 2021), waning public sentiment towards vaccines (Ali et al., 2021), effects of biased or aggressive media (Shin and Thorson, 2017), behavioral aspects of news sharing (Valenzuela et al., 2017) and information about corporate reputation (Jonkman et al., 2020). Recent research has used sentiment anlaysis for e-sports (Ardianto et al., 2020), sports (Hegde et al., 2021), patient care (Chekijian et al., 2021), healthcare apps (Camacho-Rivera et al., 2020), fear sentiment and machine learning (ML) applications (Samuel et al., 2020a; Samuel, 2017b), commodity prices (Sinha and Khandait, 2021) and even geopolitical conflicts such as the Syria Chemical attack (Bashir et al., 2021).

Recently a failed attempt to use sentiment analysis without depth of investigation meth-

Figure 1: Criticism of a Sentiment Analysis based Washington Post article on media bias

ods became famous for all the wrong reasons: Dana Milbank used sentiment analysis in his Washington Post article to claim that "*The media treats Biden as badly as ...or worse than Trump*" (Milbank, 2021). Numerous experts chimed in to point out methodological issues and absence of validity (Fig. 1), and Nate Silver tweeted "*explaining his analysis on why the " ...sentiment analysis" cited in the @milbank WaPo article this weekend is complete crap...*" and " *...Designing good algorithms is hard, but this is an especially bad one*".

1.1 Definitions and overview of SA

Sentiment analysis (SA) is being widely used across research disciplines and practitioner domains, and is often deemed to be "omnipresent as a concept" but is ridden with a number of analytical, domain-specificity, methodology and interpretation challenges (Van Atteveldt et al., 2021). There have been a broad range of definitions used for SA. SA has been defined as "a generic name for a large number of opinion and affect related tasks" (Mohammad, 2017), as "a research field that aims at understanding the underlying sentiment of unstructured content" (Poria et al., 2020), as "an active research area to display emotions and to automatically discover the sentiments expressed within the text" (Nazir et al., 2020), and as "a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language" (Mäntylä et al., 2018). In its early stages in particular, and to some extent event to present date, there appears to exist some conflicting usage of the term "opinion mining" along with SA. Many studies use the terms as synonyms without concern for their implicit meanings. For example, SA has been defined as "the gathering of people's views regarding any event happening in real life ...understanding the emotions of the people stands extremely important" (Chakraborty et al., 2020), as "the process of user's opinion extraction regarding a topic, event, entity or a situation through analyzing unstructured data from tweets (mainly the text that a tweet contains) (Britzolakis et al., 2020), and in a more accommodating way as both "opinion mining, dealing with the expression of opinions" and "emotion mining, concerned with the articulation of emotions (Yadollahi et al., 2017). Other popular approaches to SA include multimodal sentiment

analysis (Cheema et al., 2021), aspect based, contextual, sentiment reasoning, domain adaptation, sentiment aware NLG, sarcasm analysis and sentiment bias (Poria et al., 2020).

1.2 Sentiment identification

Sentiment, an expression of emotion, as evidenced in textual data has been measured in a variety of ways, including binary classification into positive and negative sentiments (Vyas and Uma, 2018), ternary classification including neutral (Bouazizi and Ohtsuki, 2017), a range of positive-neutral-negative categories (Yadollahi et al., 2017), composite continuous scores ranging from approximate limits such as -1:+1 (Rinker, 2019) or -5 to +5 (Nielsen, 2011), and sentiment classifications such as fear, surprise and joy (Jockers, 2017). Sentiment analysis has been traditionally paired with opinion mining (Poria et al., 2020).

1.3 SA: A Clear Definition

However, we believe that sentiment analysis has diverged significantly from opinion mining in goals and methods over the past decade and therefore we consider only sentiment analysis from a natural language understanding (NLU) perspective for estimating human emotions. Hence, we define sentiment analysis as being *a process using textual analytics or NLP methods to classify emotions or to measure the extent of emotions and polarities of human feelings reflected in data, generally textual data. Sentiment analysis can also be extended to other forms of data (such as text + image or audio) using multimodal and hybrid NLP strategies*. For example, we can train models to identify emotion on the basis of facial expression and corresponding text. Interestingly, in spite of much research and significant progress, there are many challenges to automated systematic SA and sentiment understanding and this study focuses on identifying the same.

1.4 Outline

The rest of this paper is organized as follows: we review extant literature, provide and overview of methods and introduce the datasets used for the analysis. We then present our analysis, results of comparing various SA methods and a discussion on the identified dark side features. We discuss potential solutions and illustrate remedial viability. We finally present our limitations, future research, critical guidelines for applied SA and conclude.

2 Literature Review

As described above and summarized in recent reviews, there exists a wide range of SA methods and tools (Poria et al., 2020; Jockers, 2017; Rinker, 2019; Vyas and Uma, 2018). SA methods can be viewed as belonging to two common strategies - one is a lexicon /dictionary /rule-based strategy and the other is a machine learning (in some form) driven strategy (Yadollahi et al., 2017; Poria et al., 2020). Human sentiment scoring and hybrid strategies have also been used for SA. Extant research has used "language and social context" for individual level SA with network effects, and it has been shown that "sentiment features of a post affect the sentiment of connected posts and the structure of the network itself" (West et al., 2014; Miller et al., 2011). Other important approaches have focused on "semantic and sentiment similarities among words" and the use of "sentiment treebanks" which consists of "fine grained sentiment labels" for "phrases in the parse trees of sentences" with Recursive Neural Tensor Networks (Maas et al., 2011; Socher et al., 2013). Human sentiment scoring and hybrid strategies are also used for SA. Many studies have compared dictionary / lexicon and machine learning based methods and have posited the usefulness of both. Recent studies have posited the superiority of sentiment classification accuracy with state of the art machine and deep learning methods and large language models However, current methods are still problematic and many unresolved issues exist in the SA discipline.

2.1 Why Consider the Dark Side of SA?

Unfettered usage of easily available of-the-shelf SA tools, packages and libraries, without

necessary validation and safeguards may often lead to diverse and contradictory results. For example, in evaluating textual responses to sports events, results based on one lexicon may lead to SA conclusions that contradict SA conclusions obtained by using another method, and it is also possible that both can be wrong. Machine learning methods can often be computationally expensive, as they need to be contextualized to training data. When trained appropriately, ML based SA tends to perform relatively better. However, errors could be rampant even with the use of complex NLP models for SA built on BERT or GPT-3 when dealing with emerging data that embody temporal linguistic changes. Given that the most prominent methods for SA are still error prone, it will be useful to acknowledge such errors and improve our understanding of shades of the dark side associated with different SA methods.

SA anomalies need to be highlighted and studied with the intent of exploring potential error classifications, which may then point to potential solutions or error flagging mechanisms, both of which would be immensely helpful for the SA discipline. Therefore, we posit the critical need to explore this "dark side" of SA, with the intent to understand the limitations of NLP driven SA and also explore if some of these limitations can be overcome.

2.2 Past concerns with SA

There have numerous studies in the recent past which have expressed concerns about the challenges with SA, and we review a few prominent studies in this section. Specifically, we review five notable studies in this section to synthesize our arguments for the dark side of SA.

2.3 SA Validity: Human, Crowd, Lexicon and ML-based approaches

Van Atteveldt et al. (2021) provide an excellent comparison of SA methods, contrasting manual annotations, crowd sourcing, lexicons and machine learning methods. They present their findings which indicate that human and crowd-sourced SA provided best performance, lexicons performed poorly and machine learning based SA performed better than lexicons but did not match the accuracy levels provide by human coded SA. Their research uses Dutch "economic headlines" for evaluation and comparison of SA methods. The study thoroughly examines the process of manual assignment of SA scores, including crowd sourcing of sentiment codes. They reviewed multiple dictionaries for this study, and employed machine and deep learning techniques. They concluded that human annotation provided the best results, along with crowd sourcing of sentiment coding. The methods using dictionaries failed to impress and deep learning techniques provided significantly better performance than dictionary-based methods. They concluded by emphasizing the importance of matching methods to the tasks and verifying the validity of automated SA methods before concluding any research based on SA (Van Atteveldt et al., 2021).

2.4 Common Mistakes in SA and Silver Bullets

In their financial SA study, (Xing et al., 2020) state that the objective of financial sentiment analysis (FSA) is to " classify a piece of financial text as expressing bullish or bearish opinions toward certain arguments". The authors identify identify six important error dimensions which have in-principle lessons for domains beyond finance: *"Irrealis Moods"*: counterfactual moods - can imply alternatives, Subjunctive mood - can carry a sense of mized sentiment, Imperative mood - an emphatic note; *Rhetoric*: can imply negative assertion; *dependent opinion*; *unspecified Aspects*; *unrecognized Words*: entity, micro text, jargon; and external references. They conclude by suggesting a mixed methods approach to tackling these unusual SA challenges in FSA (Xing et al., 2020).

2.5 SA Domain: Mature or Not?

In their study on present challenges and the future directions in SA research, (Poria et al., 2020) provide a comprehensive review of the

backdrop of SA research over the decades, and present clear frameworks summarizing developments in this broad discipline. They posit that "there is an underlying perception that this field has reached its maturity" and then go on highlight the "shortcomings and underexplored, yet key aspects of this field necessary to attain true sentiment understanding". They then analyzed the major developments that popularized SA and charted potential avenues for future development of the and also raised "many overlooked and unanswered questions" (Poria et al., 2020). They identified the following challenges, some of which are commonly known, and yet the authors provided additional insights on these: Lexicons are good with words and phrases but can fail with sentences, especially long and complex sentences, and also fail due to inability to account for context, and subjectivity in annotations; SA for dialogues versus monologues; SA in varying or mixed cultural contexts; SA in the presence of sarcasm; and SA for creative language usage. This study is a significant motivator for the present research as it lays the ground work for a clearer articulation of the dark side of sentiment analysis.

2.6 Ethical Challenges in SA

Mohammad (2017) specifically highlights issues with SA of "a combination of terms" and phrases which include "negators, degree adverbs, and intensifiers". His chapter on challenges with SA emphasize that even though many SA methods are ridden with limited accuracy issues, yet they "accurately capture significant changes in the proportion of instances that are positive (or negative)". Mohammad (2021) also address the ethical challenges of SA in affective computing and the use of "automated emotion recognition" (AER) mechanisms. The study elaborates on fifty ethical considerations clustered into groups by "Task Design, Data, Method, Impact and Evaluation, and Implications for Privacy and Social Groups".

2.7 Aspect, multimodality & label errors in SA

Nazir et al. (2020) address the issues surrounding aspect based sentiment analysis, an emerging dimension of SA which aims to identify and extract aspects in text, analyze sentiment and study sentiment evolution (SE). The complexity of aspect based SA and SE are rooted in the challenges of aspect identification and extraction, cross-domain transfer learning, multimodal data, context specific semantics, and temporal sentiment and emotional dynamism. Cheema et al. (2021) conduct a multimodal SA study, which is a fascinating and emerging area of SA. Under this method, multiple data formats such as text and images are analyzed to create a composite SA measure reflecting the underlying emotion. For example, social media posts may contain images with embedded text or images with a caption or text supported by images. Multimodal SA can be challenged by components with conflicting sentiment implications - this means that in addition to all the challenges of SA for textual data, there could be images associated with the text that can imply opposing sentiment or tangential sentiment or plurality of sentiment than that which is indicated by text alone. The positive aspect is that multimodal SA comes closer to real world emotion detection and if developed and validated, can lead to holistic automated SA. Another interesting challenge to SA comes in the form of label errors, which can be a result of subjective human labeling in lexicons, faulty crowd-sourcing or poor quality train, dev or test data. Northcutt et al. (2021) analysed 10 commonly used benchmark datasets for a broad range of AI technologies such as NLP, computer vision and sound /speech recognition and demonstrated label errors, and such label errors have a the likelihood of creating model errors leading to faulty SA.

2.8 Common SA Packages in R

Aside from sentimentr, one of the most commonly used packages for performing SA in R is the Syuzhet package (Jockers, 2015). Like sentimentr, Syuzhet provides out-of-the-box

support for several common lexicons, but the common use of package with out-of-the-box default options highlights some of the perils of blinding deploying SA tools without understanding the context or domain that a particular SA package was intended to serve.

Syzuhet's default settings for calculating sentiment scores are to use a custom sentiment dictionary that was created from a corpus of fiction novels. While such a dictionary would be appropriate in the context of studying fiction literature, this type of lexicon has been identified as having shortcomings for other types textual data. SA users need to carefully consider the limitations of lexicons before deploying them in light of the limitations present and domain-specific intended applications. Syuzhet also supports some of the other more commonly deployed sentiment dictionaries.

Naldi (2019) compared Syuzhet with sentimentr and two other commonly used R packages and identified some limitations. Syuzhet does not provide support for customizing sentiment dictionaries. The lack of customizability potentially opens up.

3 Data, methods, analysis and results

We are interested in demonstrating that SA is not a straightforward slam-dunk tool-driven process, and that there are significant 'dark side' challenges to applying automated SA with lexicons and dictionaries based methods. Therefore, in addition to the literature review based identification of open problems, this study focuses on exploring the dark side of SA using Twitter data to explore the following:

1. Assign sentiment scores to a collection of tweets
2. Compare multiple lexicons and dictionaries based sentiment scores
3. Compare multiple lexicons and dictionaries based sentiment classes
4. Identify some of the challenges for SA with dictionary and lexicon based approaches
5. Discuss SA using exploratory and accuracy evaluation methods
6. Highlight fallacies and the "Dark Side" of SA
7. Identify opportunities to mitigate the dark side of SA

3.1 Data and Lexicon-based SA

For the initial part of our data analysis, we used the R statistical programming language. We used the R package "sentimentr" for applying the lexicons and obtaining sentiment scores. These scores were then converted to positive, neutral and negative classes. We used the default polarity and valence shifter settings with an of-the-shelf approach to applying SA to the data. The following lexicons were used: Jockers (V1), Jockers-Rinker (V2), Loughran-Mcdonald (V3), NRC (V4), Senticnet (V5), Huliu (V6), and Socal-google (V7) [Figure 2]. Our dataset for the initial exploratory analysis consisted of 5789 tweets, randomly selected from a collection of 2021 tweets with "market" and "vaccine" as keywords (Samuel et al., 2021). Since the amount of vaccine tweets were disproportionately higher in the raw data, stratified random sampling was used to ensure a balance of items in the first dataset which was finally composed up of 2895 "market" tweets and 2894 "vaccine" tweets.

3.2 SA measure: Variance, Correlation and Confusion Matrix

We start our data analysis by qualitatively observing a high variance across classes - that represents a high level of fluctuation in the sentiment scores assigned by multiple lexicons for the same tweet, as shown for the plots of sentiment scores for 5789 tweets in figures 2 and 3. Secondly, we discuss and plot correlations among the various lexicons. We emphasize that correlations do not serve as a measure of SA accuracy - rather we simply use correlations to explore and demonstrate the lack

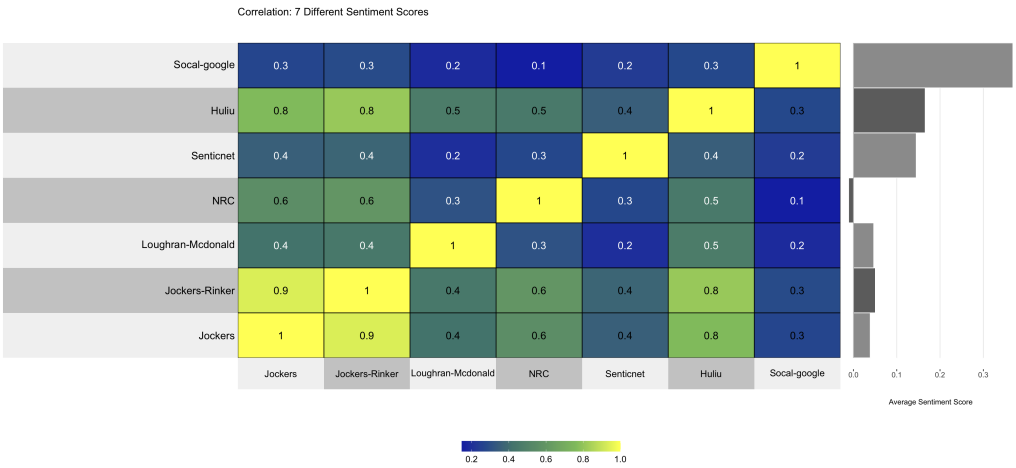


Figure 2: SA classifications for 5789 tweets with correlations

of agreement between SA lexicons and dictionaries. An absence of a high correlation between sentiment scores of multiple lexicons indicates that the sentiment scores are moving without significant agreement of scores against a given set of texts. The existence of such low correlations among SA Lexicon scores serve as a somber warning against any of-the-shelf application of SA lexicons and dictionaries without additional method logical support and validation of accuracy. For the purposes of determining accuracy we use confusion matrices, where the actual score is based on human expert classification ("gold" standard) and predicted scores are given by SA lexicons as visualized in figures 4 and 5, and summarized in table 1. A confusion matrix is a presentation of model classification output which represents accuracy by contrasting actual versus predicted classes, and is widely used in "machine learning to evaluate the quality of a classifier" by cross-classifying "predicted and actual decision classes" and is also known as the "error matrix" (Düntsch and Gediga, 2019).

3.3 Variance in dictionary / lexicon-based SA

We observed a high measure of disagreement in sentiment scores and classes among the lexicons in their sentiment scores for the same 5789 tweets. Figure 2 demonstrates this phenomenon, and though there is a fair correlation of 0.93 between Jockers and Jockers-Rinker (V1 and V2), all the other lexicons fail to provide any significant measure of agreement in a plain vanilla of-the-shelf application. We did not create a "gold standard" sentiment score for all the 5789 tweets and hence we do not comment on the intrinsic accuracy of any of the lexicons in this part of our analysis, but it suffices to know that this level of disagreement indicates significant challenges to any direct of-the-shelf application of lexicons to generate meaningful sentiment scores or classes. We observed, not surprisingly from a mathematical perspective, that the mean variance by sentiment class was a very high 0.41 as compared to a mean variance by sentiment score of 0.10. This simply implies that the raw scores provided by lexicons may contain positive or negative scores close to the neutral score of 0, qualifying the degree to which the sentiment is positive or negative. This can sometimes be sufficiently informational as compared to more rigid binary or ternary or quaternary forms of sentiment classification.

3.4 SA: Lexicons and Gold standard

For a more detailed analysis and review of SA accuracy, we created a subset of randomly selected 399 market and vaccine tweets (as time and resources permitted) with a manually annotated "Gold standard" sentiment classification into positive, neutral and negative classes. Our

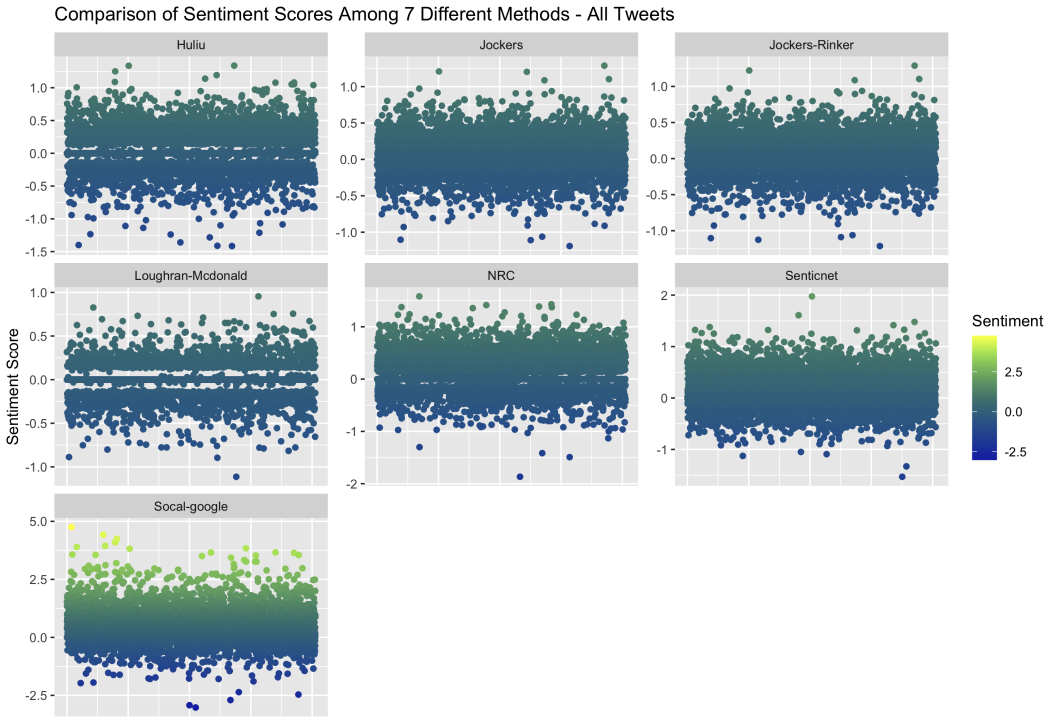


Figure 3: Scatter plot comparison of SA classifications for 5789 tweets.

comparison using of-the-shelf applications of seven sentiment lexicons on 399 tweets with the human scored Gold standard and presented in a visual summary form in figure 4. The results of this comparative analysis reflect the **dark** side of SA [Figure 4] and are summarized in table 1.

We applied the nonparametric Spearman correlation and identified that the highest correlation between Gold and any of the lexicons approached 0.3 and the lowest correlation was a measly 0.13 [Figure 4]. This implies alarmingly high levels of disagreement among of-the-shelf lexicon based SA performance! We discuss potential reasons for this **dark** SA performance by lexicons in the discussion section.

3.5 SA: Lexicon Calibration & Custom Dictionaries

The sentiment scores for this study are obtained by an of-the-shelf approach with applying lexicons for SA. However, it is possible to improve the accuracy of SA by calibrating the lexicon sentiment scoring settings. For example, additional insights can be obtained using NRC's Valence, Arousal, and Dominance (VAD) Lexicon which lists over 20,000 English words along with their corresponding "valence, arousal, and dominance scores" (Mohammad, 2018). An important work in this direction has been implemented by the `sentimentr` package in R, which attempts to interpret and score the influence of "valence shifters" (Rinker, 2019). Simple stated, valence shifters refer to negations and adversarial arrangements of words in sentences which tend to flip the polarity of the underlying sentiment associated with a word or a set of words. Furthermore, it is possible to create custom lexicons from scratch and such lexicons can lead to greater accuracy with a domain or discipline specific approach to customization. It is also possible to customize SA dictionaries with sentimentr which supports "making and updating" or dictionaries by polarity or valence shifter (Rinker, 2019).

3.6 The Statistical Monkey & Chimp Perform SA

This comparison also includes a best of "Monkey" classification with a purely random equal chance assignment of sentiment classes using

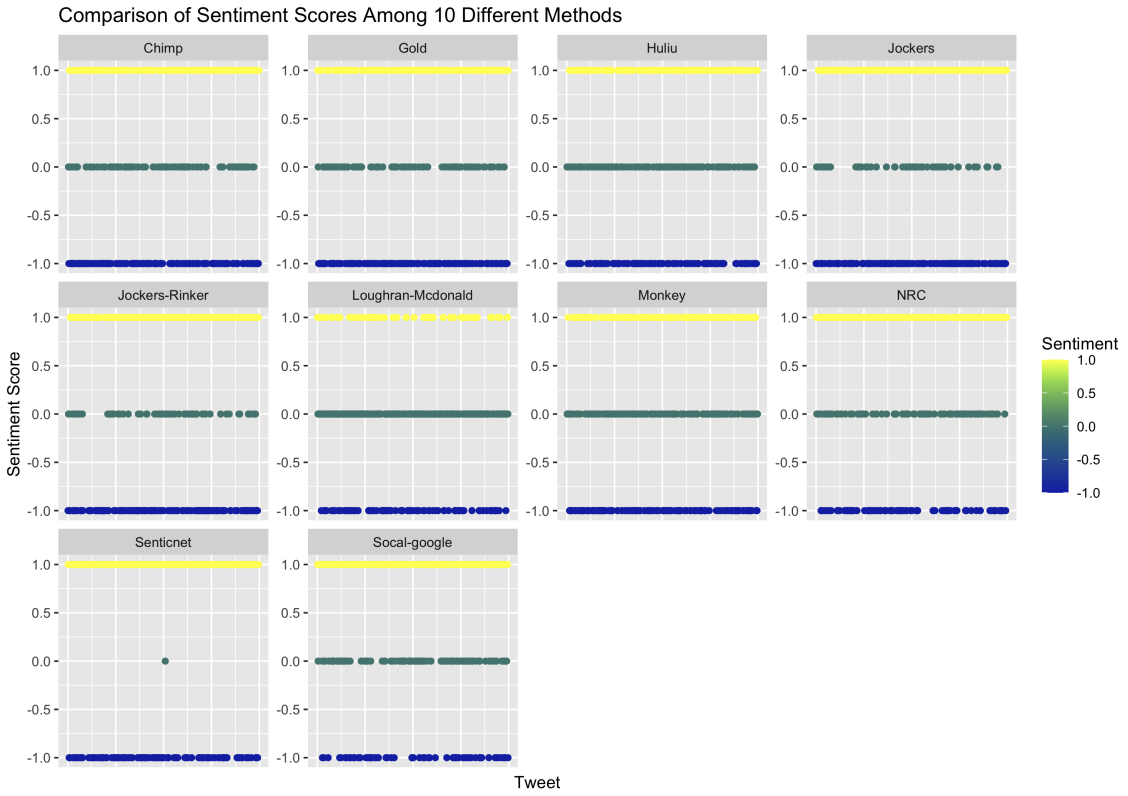


Figure 4: Comparison of SA classifications & Gold standard for 399 tweets- plots, distributions & correlations.

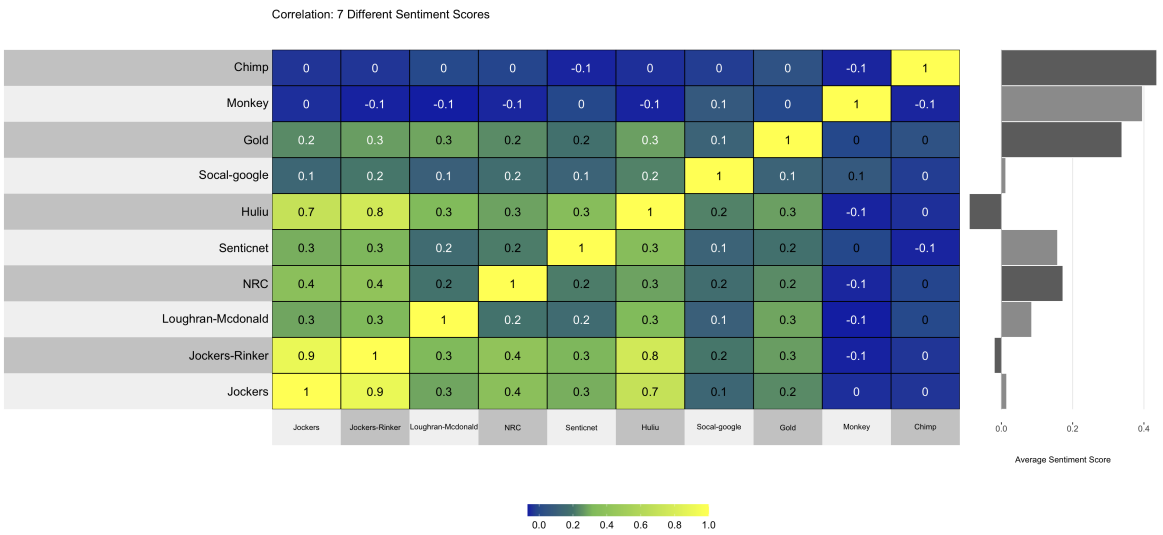


Figure 5: Comparison of SA classifications: Lexicons, Gold, Generic-ML & Random (Monkey & Chimp)

a pseudorandom number generation process, and a best of "Chimp" classification with a weighted (positive = 152, neutral = 87, negative =160, total =399) random assignment of sentiment classes, as summarized in table 1. We remind readers that this is a generic random integer (-1, 0 and 1) based classification, and not developed with data specific rules. Most of the top lexicons are beating the Monkey and the Chimp. The Monkey performed the worst of all SA methods but almost matched Loughran-Mcdonald, but the Chimp

performed better than Loughran-Mcdonald and Socal-google as summarized in Table 1 using confusion matrices for all 9 methods.

4 Throwing Light on the Dark Side

Though the present study has performed direct comparisons of SA scores from multiple dictionaries / lexicons to ensure parsimony in an exploratory study, it will be necessary for future research to study the underlying methods to ensure statistically valid comparisons. For example, subject to the items being studied, it may be necessary to apply normalization / standardization of SA scores or ensure data specific customization to obtain sentiment scores.

Our analysis identified a few reasons for poor performance of lexicons and some ideas which could improve lexicon performance. None of the suggestions by themselves should be expected to necessarily lead to near-perfect or improved solutions, but we have seen improvements in our experiments by implementing these on subsets of of our experimental data:

- Context - create and use context specific lexicons.
- Study the scoring systems used by each dictionary or lexicon, and adapt SA processes to the specific methods being used.
- Spelling errors - correct bad spellings or expand the dictionary /lexicon to accommodate errors.
- Mixed emotions - these are more difficult to address with lexicons and ideas from aspect based SA would be useful.
- Valence shifters - some arrangements of words, as evidenced in adversarial examples, may escape standard valence shift detection and hence it will be useful to make necessary adjustments based on nuances of the data being processed
- Twitter-speak challenges - It will be useful to create data-source-specific and topic-specific sentiment lexicons, and extend that in principle to platform or source specific lexicons.
- Use alternative approaches: apply human hand coding where possible or to the extent possible to validate - for example, it may be preferable to avoid automation for SA with small or complex data.
- Improve input data - sourcing high quality and relevant data is a critical part of the SA pipeline. For example, reduction of ambiguity, acronyms and equivocal words and phrases would help improve the quality of input data.

4.1 SA Dictionaries OR Lexicons?

"Dictionary" and "lexicon" are two distinct words with individual meanings. Yet, we find some ambiguity in their usage in NLP, specifically in the context of sentiment analysis where the two words appear to be often interchangeably used as "Sentiment Dictionary" and "Sentiment Lexicon". In developing a sentiment dictionary for slang words, (Wu et al., 2018) use dictionary and lexicons fluidly: *"To this end, we propose a web-search-based, learning approach to build the first slang sentiment word dictionary, named SlangSD, ...and it (SlangSD) can be easily incorporated as an additional sentiment lexicon."* (Wu et al., 2018). Similarly, it has been considered fair to compare lexicons and dictionaries and similar purpose instruments: *"Compared with other lexicons, the dictionary generated using our approach is language-independent..."* (Rao et al., 2014). It has also been observed that well known self-declared lexicons have also been called dictionaries: The authors of SENTIWORDNET called it *"A Publicly Available Lexical Resource for Opinion Mining"* and the word "dictionary" is not found in their seminal paper (Esuli and Sebastiani, 2006). However, subsequent research has called it a dictionary: *" ...have developed the SentiWordNet dictionary based on WordNet dictionary"* (Ameur and Jamoussi, 2013). In some cases, authors qualify the progression from one word to the

Accuracy	CV1	CV2	CV3	CV4	CV5	CV6	CV7	Mnky	Chimp
Correct	195	201	151	179	185	180	160	148	167
Incorrect	204	198	248	220	214	219	239	251	232
% Correct	49%	50%	38%	45%	46%	45%	40%	37%	42%

Table 1: Summary of Classification Accuracy: Lexicons & Dictionaries, Monkey & Chimp for SA

other without providing a reason for the shift: *"Hereafter, we will call these standardized sentiment lexicons as sentiment dictionaries"* (Cho et al., 2014). We also found a creative phrase with "lexicon dictionary": "...extractScore() uses the lexicon dictionary L to associate each word..." (Ahmed et al., 2020).

4.1.1 SA Dictionaries and Lexicons!

We propose that the distinction between "Dictionary" and "lexicon" be respected in SA as two distinct words associated with their individual meanings applied to the SA discipline. We suggest a few early stage thoughts towards this:

- Lexicons should list words, phrases or character-sets (slangs, emoticons) under classes (positive, negative, fear, joy, etc; and where required, associated integer values). Artifacts and systems which fall into this category can be called lexicons.
- Dictionaries should list words, phrases or character-sets (slangs, emoticons) and provide methodological arrays of associated words, phrases or character-sets and where required, scaled or relative quantitative values as expressions of some form of sentiment-meaning. Artifacts and systems which fall into this category can be called dictionaries.
- In a lighter vein, but for a valid point, when artifacts and systems combine features of both the above categories, then such a "lexicon dictionary" entity may be called a *"lexionary"*, or perhaps a *"dixicon"*?.

In spite of the widespread interchangeable past usage, we suggest that adopting some such nomenclature will provide a better framework for future research. Else, this will serve as a example where NLP research contributes to the dark side by increasing undesirable linguistic equivocality.

4.2 Future Work: The Dark side of SA

This study is part of a research project on challenges in natural language understanding (NLU) and the next part addresses the dark side of machine learning based SA. Future research on SA could also consider adopting non-text variables such as has been used fir modeling social media virality (Garvey et al., 2021). While this study successfully demonstrates the dark side of SA, much work remains to be done to develop frameworks to evaluate SA applications and output, and identify the key discrepancies or measures reflecting the presence of inaccuracies. This is important because the consequences of the dark side of SA are significant. For example, shallow application of of-the-shelf SA tools can result in:

- Rejection of good product features based on false -ve sentiment scores
- Errors in SA can lead to wrong political expectations
- SA false flags can enhance news based polarization (See fig. 1)
- SA based processes and systems such as sentiment sensitive dialogue applications can experience chronic failures due to the dark side of SA

This study makes a notable cautionary contribution: it identifies the significant variation of SA output for the same data with multiple lexicons, and highlighted that most of the SA lexicons and dictionaries failed to perform

satisfactorily, with some even being matched by random selection processes (monkey and chimp).

Our review indicates that there is significant scope for improvement in SA methods (one size does not fit all scenarios /contexts), SA datasets and SA tools and applications. Increasing interest in SA in applied research and practice demonstrate the need to pursue further development of the SA domain. We plan to expand the Gold data with more human expert-coded labels and study the dark side of SA with machine learning methods, including Transformers.

This study is part of a series of NLU studies, and the next studies which are underway include:

- The dark side of SA with custom sentiment lexicons, dictionaries and rules.
- The dark side of SA with machine learning and language models
- The bright side of SA - hybrid approaches to SA

5 Conclusion

"There are a number of good fixed lexicons for sentiment. They [show] negligible to high levels of disagreement with each other. These can be exploited strategically — resolve the conflicts somehow or allow them to persist as genuine points of uncertainty."
- Christopher Potts, Stanford.

The above quote posits some important aspects of pragmatic SA implementation: lexicons and dictionaries help in SA, in-depth analysis of SA is necessary before drawing conclusions, it is important to know the limits of SA methods and tools and SA modeling may need to be customized for some situations, while it is better to acknowledge the absence of satisfactory SA solutions for some situations. SA tools are very useful and must continue to be used for research and practice - however, as cautioned, it is vital to understand the conflicts and ways to acknowledge and address them. This study discussed known issues with SA as documented by prior research and then compared the application of multiple of-the-shelf SA lexicons and dictionaries and Monkey and Chimp random number methods to stock market and vaccine tweets. It was not our intent to improve accuracy, but rather to highlight the dark side of SA by highlighting SA discrepancies and spur a crucial discussion on the characteristics of the dark side of SA. This research will help align researcher and practitioner expectations to carefully consider the known limitations and boundaries of presently available of-the-shelf SA lexical tools and methods. We hope that this study will lead to deeper attention to applied SA and spur new strategies for the improvement of sentiment analysis research and practice.

Acknowledgement

- This study was initiated as part of the Artificial Intelligence professional program at Stanford University. We acknowledge the valuable guidance from Prof. Chris Potts, and from faculty and staff at the Stanford Center for Professional Development, and in particular, Samir Sen and Steve Haraguchi.
- This study was supported by the RUCI lab at Rutgers University https://rucilab.rutgers.edu/about-ruci/)
- Limitations or errors, if any, are our own.

References

Murtadha Ahmed, Qun Chen, and Zhanhuai Li. 2020. Constructing domain-dependent sentiment dictionary for sentiment analysis. *Neural Computing & Applications*, 32(18).

GG Ali, Md Mokhlesur Rahman, Amjad Hossain, Shahinoor Rahman, Kamal Chandra Paul, Jean-Claude Thill, Jim Samuel, et al. 2021. Public perceptions about covid-19 vaccines: Policy implications from us spatiotemporal sentiment analytics. *Healthcare*.

Hanen Ameur and Salma Jamoussi. 2013. Dynamic construction of dictionaries for sentiment classification. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 896–903. IEEE.

Rian Ardianto, Tri Rivanie, Yuris Alkhalifi, Fitra Septia Nugraha, and Windu Gata. 2020. Sentiment analysis on e-sports for education curriculum using naive bayes and support vector machine. *Jurnal Ilmu Komputer dan Informasi*, 13(2):109–122.

Saimah Bashir, Shohar Bano, Sheikh Shueb, Sumeer Gul, Aasif Ahmad Mir, Romisa Ashraf, Neelofar Noor, et al. 2021. Twitter chirps for syrian people: Sentiment analysis of tweets related to syria chemical attack. *International Journal of Disaster Risk Reduction*, 62:102397.

Mondher Bouazizi and Tomoaki Ohtsuki. 2017. A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access*, 5:20617–20639.

Alexandros Britzolakis, Haridimos Kondylakis, and Nikolaos Papadakis. 2020. A review on lexicon-based and machine learning political sentiment analysis using tweets. *International Journal of Semantic Computing*, 14(04):517–563.

Marlene Camacho-Rivera, Huy Vo, Xueqi Huang, Julia Lau, Adeola Lawal, and Akira Kawaguchi. 2020. Evaluating asthma mobile apps to improve asthma self-management: user ratings and sentiment analysis of publicly available apps. *JMIR mHealth and uHealth*, 8(10):e15076.

Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. 2020. Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, 97:106754.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. *CoRR*, abs/2106.08829.

Sharon Chekijian, Huan Li, and Samah Fodeh. 2021. Emergency care and the patient experience: Using sentiment analysis and topic modeling to understand the impact of the covid-19 pandemic. *Health and Technology*, pages 1–10.

Heeryon Cho, Songkuk Kim, Jongseo Lee, and Jong-Seok Lee. 2014. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems*, 71:61–71.

Ivo Düntsch and Günther Gediga. 2019. Confusion matrices and rough set data analysis. In *Journal of Physics: Conference Series*, volume 1229, page 012055. IOP Publishing.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Myles D Garvey, Jim Samuel, and Alexander Pelaez. 2021. Would you please like my tweet?! an artificially intelligent, generative probabilistic, and econometric based system design for popularity-driven tweet content generation. *Decision Support Systems*, 144:113497.

Siddhanth U Hegde, AS Zaiba, Y Nagaraju, et al. 2021. Hybrid cnn-lstm model with glove word vector for sentiment analysis on football specific tweets. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–8. IEEE.

Matthew Jockers. 2017. Package 'syuzhet'. *URL: https://cran. r-project. org/web/packages/syuzhet*.

Matthew L. Jockers. 2015. *Syuzhet: Extract Sentiment and Plot Arcs from Text*.

Jeroen GF Jonkman, Mark Boukes, Rens Vliegenthart, and Piet Verhoeven. 2020. Buffering negative news: Individual-level effects of company visibility, tone, and pre-existing attitudes on corporate reputation. *Mass Communication and Society*, 23(2):272–296.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

Dana Milbank. 2021. Opinion — Biden's media coverage is worse than Trump's at times - The Washington Post.

Mahalia Miller, Conal Sathi, Daniel Wiesenthal, Jure Leskovec, and Christopher Potts. 2011. Sentiment flow through hyperlink networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Saif M Mohammad. 2017. Challenges in sentiment analysis. In *A practical guide to sentiment analysis*, pages 61–83. Springer.

Saif M Mohammad. 2021. Ethics sheet for automatic emotion recognition and sentiment analysis. *arXiv preprint arXiv:2109.08256*.

Ashley Muddiman and Natalie Jomini Stroud. 2017. News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication*, 67(4):586–609.

Maurizio Naldi. 2019. A review of sentiment computation methods with r packages.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: a comprehensive survey. *IEEE Transactions on Affective Computing*.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.

Alexander Pelaez, Elaine R Winston, and Jim Samuel. 2021. David and goliath revisited: How small investors are changing the landscape of financial markets. *Northeast Decision Sciences Institute (NEDSI)*, 2021-50:287–292.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

Md Mokhlesur Rahman, GG Md Nawaz Ali, Xue Jun Li, Jim Samuel, Kamal Chandra Paul, Peter HJ Chong, and Michael Yakubov. 2021. Socioeconomic factors analysis for covid-19 us reopening sentiment with twitter and census data. *Heliyon*, 7(2):e06200.

Yanghui Rao, Jingsheng Lei, Liu Wenyin, Qing Li, and Mingliang Chen. 2014. Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17(4):723–742.

Tyler W. Rinker. 2019. *sentimentr: Calculate Text Polarity Sentiment*. Buffalo, New York. Version 2.7.1.

Jim Samuel. 2017a. Informatics in information richness: A market mover? an examination of information richness in electronic markets. *Samuel, J. & Pelaez, A.,(2017). Informatics in Information Richness: A Market Mover*.

Jim Samuel. 2017b. Information token driven machine learning for electronic markets: Performance effects in behavioral financial big data analytics. *JISTEM-Journal of Information Systems and Technology Management*, 14(3):371–383.

Jim Samuel, GG Ali, Md Rahman, Ek Esawi, Yana Samuel, et al. 2020a. Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, 11(6):314.

Jim Samuel, Ratnakar Palle, and Eduardo Correa Soares. 2021. Textual data distributions: Kullback leibler textual distributions contrasts on gpt-2 generated texts, with supervised, unsupervised learning on vaccine & market topics & sentiment.

Jim Samuel, Md Mokhlesur Rahman, GG Md Nawaz Ali, Yana Samuel, Alexander Pelaez, Peter Han Joo Chong, and Michael Yakubov. 2020b. Feeling positive about reopening? new normal scenarios from covid-19 us reopen sentiment analytics. *IEEE Access*, 8:142173–142190.

Google Scholar. 2021. Sentiment — analysis.

Jieun Shin and Kjerstin Thorson. 2017. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, 67(2):233–255.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Future of Information and Communication Conference*, pages 589–601. Springer.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Sebastián Valenzuela, Martina Piña, and Josefina Ramírez. 2017. Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of communication*, 67(5):803–826.

Wouter Van Atteveldt, Mariken ACG van der Velden, and Mark Boukes. 2021. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140.

Vishal Vyas and V Uma. 2018. An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner. *Procedia Computer Science*, 125:329–335.

Robert West, Hristo S Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310.

Liang Wu, Fred Morstatter, and Huan Liu. 2018. Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52(3):839–852.

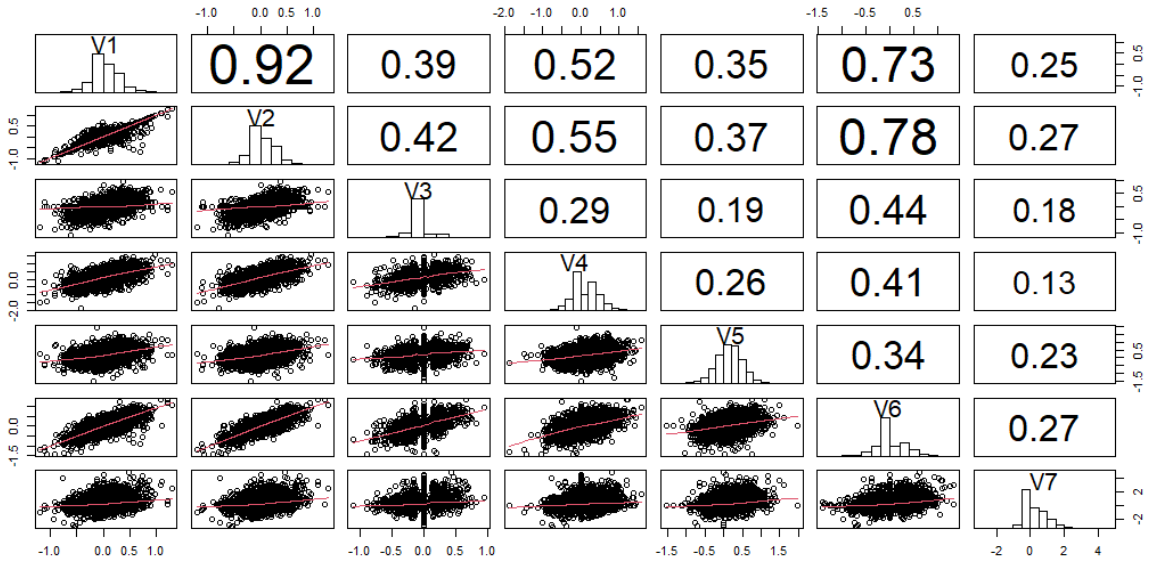
Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 978–987.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

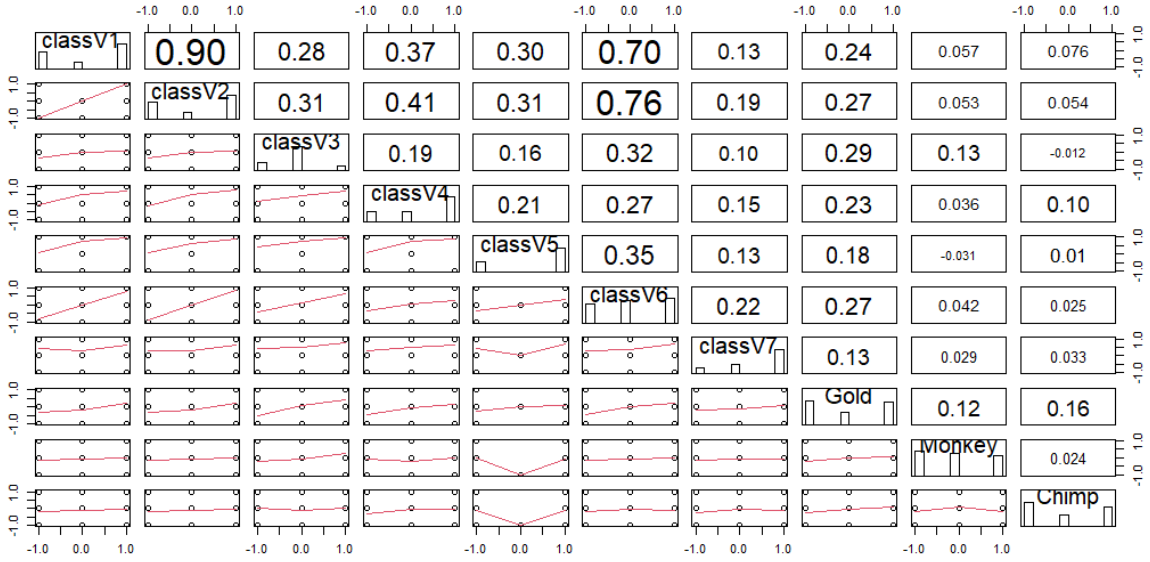
APPENDIX

Next page...

Alternate Visualizations 1 & 2



1: SA scores on 5789 tweets- scatter plots, distributions and correlations.



2: SA classifications & Gold standard for 399 tweets- plots, distributions & correlations.