Beyond VADER: A Specialized Approach to Sentiment Analysis in YouTube Live Chats

1st William DeBlieck
Dept. Mathematics and Statistics
Cornell College
Mount Vernon, USA
wdeblieck24@cornellcollege.edu

***Abstract*—Non domain-specific sentiment analysis models tend to preform poorly in environments where syntax and grammar differ from the norm. This paper presents an approach to creating a PN-polarity sentiment analysis model based on POS and punctuation and a lexicon for YouTube Live chat feeds, combining manual POS labeling and semi-automated lexicon expansion based on SentiWordNet to create a sentiment analysis model to deal with novel vocabulary. This model is then compared to Hutto's VADER model a widely used model in lexicon based sentiment analysis. The VADER model shows for the YouTube Live chat data an R^2 of 0.191 and the model built in this paper shows an R^2 of 0.639.**

***Index Terms*—Sentiment Analysis, Natural Language Processing, VADER**

I. Introduction

Most available sentiment analysis models and lexicons are built for general use. While this is usable for most sentiment analysis projects, this makes it much harder to analyze the sentiment of niche content like internet discourse. Specific models or lexicons must be built for accurate analysis of domain specific syntax. This, however, is not the approach many sentiment analysis projects with domain specific syntax are built upon[1][2][3][4]. The content that this paper tries to analyze is the YouTube Live chat feed. YouTube live is a live streaming platform in which a person may share real time video content with other users and interact with users watching through a chat function. The live chat feature on YouTube hosts a distinct sub-dialect within the realm of internet English. This papers approach involves creating a new lexicon model, built upon an expanded version of SentiWordNet, a widely-used lexical resource for sentiment analysis. The new model and lexicon will then be compared to the performance with a general social media lexical model: VADER[9] with the prebuilt lexicon from the nltk library. VADER is a widely used sentiment analysis lexicon model. It has been used on many projects from research on the website 4Chan's opinion on CoVid-19 to stock market advice for what stock is most positively talked about on Reddit[4][3]. Unlike in Kim et. al, Hu, and Goldensohn et. al's papers[5][6][7] we are not categorically using sentiment topic catagories with this paper's sentiment analysis. This is due to the nature of live chat feeds in which the topics change greatly from one second to the next. While topic categorization may have some use in very large live chat feeds it would not have any use in the majority of live chat feeds.

II. Methodology

To prepare the model a multistep process was used. The first step was the data collection an initial cleaning. this entailed getting the YouTube Live chat data and turning it into a form useable for the lexicon building. The next step was the lexicon building. This entailed building off of a prebuilt lexicon to build a lexicon that includes the domain specific syntax and transforming the lexicon into a usable form for the model. The final step was processing the message data to make the variables used in the model.

A. Data Collection and Processing

The primary method of data collection was through the use of a modified pre-built web scraper[13], which recorded the chat feeds of a select number of past YouTube Live. The web scraper recorded all chat messages, along with the corresponding timestamps, usernames, and other metadata.

Data was collected from a handpicked selection of 4 recorded YouTube Lives, capturing the real-time interactions between the streamer and the audience. The training data consisted of a message randomly selected from each of YouTube Live picked, along with 250 to 700 subsequent messages.

The data collection and processing also included a manual labeling process, where each chat message was labeled by the researcher as either "Very Positive", "Positive", "Neutral", "Negative", or "Very Negative", based on its overall sentiment. This manual labeling process was necessary due to the dynamic and context-dependent nature of sentiment in YouTube live chat feeds.

After data collection, the research design involved a series of data preprocessing steps. The first step was the removal of all stop words from the live chat data. This was done by cross checking the words in the chat feed messages with the nltk stopword word bank. Next the final punctuation mark (defining punctuation as ".", ",", "?", or "!") was saved for each message before punctuation was removed from each message. This was done so that final punctuation could be tested as a predictor for sentiment. Punctuation can be used as a surrogate for aspect type (implicit or explicit) and for the severity of the sentiment expressed. Aspect type has a history of use as a predictor for sentiment[12].

TABLE I
SENTENCE TYPE AND EMOTIONAL LEVEL

Base Sentence	**Aspect Type**	**PN-Polarity**
”You are going to the party.”	Explicit	Weak
”You are going to the party?”	Implicit	Weak
”You are going to the party!”	Explicit	Strong

The use of a ”.” or ”?” can change a sentence’s aspect as seen in Table 1. An ”!” can increase the PN-polarity as discussed in Oad’s paper[10].

Emojis were removed from the live chat messages. This is due to there being hundreds of thousands of emojis used on YouTube. Each creator can have their own emojis and they can have different meanings within each creator’s community.

B. Lexicon Building

The lexicon built for the model discussed in this paper is an extended SentiWordNet with SO-polarity removed and new terms, sentiment, and parts of speech added. This is first done by making a surrogate compound polarity variable and replacing all polarity variables in SentiWordNet with it. Then all words, after preprocessing of the YouTube Live feed data, that are not in SentiWordNet are added and marked with average sentiment and part of speech.

SentiWordNet tracks neutral, positive, and negative polarity for a large lexicon. As SO-polarity would complicate the data annotation process all polarity scores were made into a compound polarity to transform them into a usable form as PN-polarity. Compound scores are generally found with the normalization of positive and negative polarity within the data [11]. Since all training data for SentiWordNet is not available this is not possible. A surrogate compound polarity score is used instead where positive and negative polarity are summed and divided by two. Using this allows us to keep the PN-polarity while removing the SO-polarity. This allows SentiWordNet to be congruent with this studies data annotation process.

SentiWordNet is then crosschecked with all words in the YouTube Live chat feed data post preprocessing. If the word is not found in SentiWordNet it is added to the new lexicon. All new words are assigned a compound sentiment polarity by the average sentiment, given by the manual data annotation, of the messages they were in. Part of speech is then assigned to them manually in the SentiWordNet form of the tags being: noun, verb, adverb, adjective, or other.

C. Processing of Messages

The message data used in the exploratory data analysis for this paper is average word compound sentiment, average noun compound sentiment, average adverb compound sentiment, average adjective compound sentiment, average verb compound sentiment, last punctuation mark, and the individual counts of all punctuation marks.

III. RESULTS

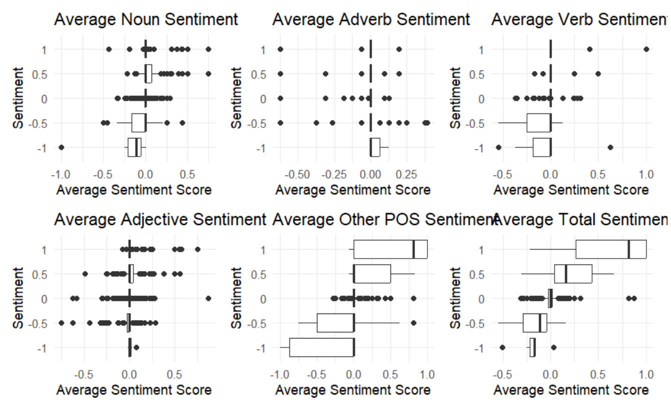

Fig. 1. Part of Speech .

When looking at sentiment in the POS, as shown in figure 1, it is apparent that average total sentiment based on the lexicon is most heavily associated with the the the actual sentiment out of all POS. It also can be seen that adverbs have little impact on the sentiment and are almost always of neutral sentiment. average other POS and average total sentiment seem to both associate in the same form as each other on all levels except for extremely negative, but this may be due to the low number of extremely negative data points. Nouns, verbs, and adjectives seem to have some association with actual sentiment.

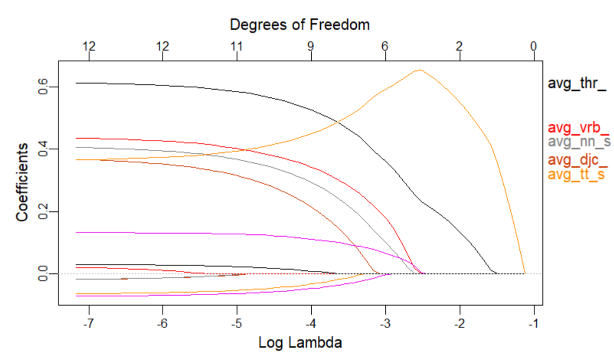

Fig. 2. Log Lambda vs Coef for Lasso Reg.

From figure 2 can see that none of the punctuation counts are good predictors as they zero out at between 9 and 6 degrees of freedom. The five quantitative predictors found are average sentiment of the words that had a POS tag of other chat message from lexicon, average sentiment of the total word sentiment of the chat message from the lexicon, average sentiment of the words with a POS tag of verbs from the chat message from the lexicon, and the average sentiment of the words with a POS tag of adjectives from the chat message from the lexicon.

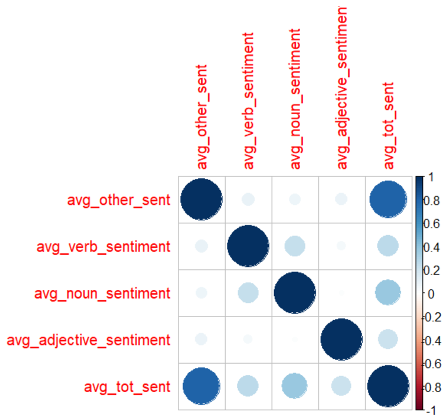

Fig. 3. Correlation plot.

Running a correlation shows an independence issue with the average sentiment of the words marked other in the POS tagging and the average word sentiment of the whole sentence. No other problems of independence are found. Due to this it must be found which predictor would be better of the two in a model containing all other current predictors.

The linear model with predictors average sentiment of words in live chat message with a POS tag of other, average sentiment of words in live chat message with a POS tag of verb, average sentiment of words in live chat message with a POS tag of noun, and average sentiment of words in live chat message with a POS tag of adjective as predictors has an R^2 of 0.6108 while The linear model with predictors average sentiment of words in live chat message, average sentiment of words in live chat message with a POS tag of verb, average sentiment of words in live chat message with a POS tag of noun, and average sentiment of words in live chat message with a POS tag of adjective as predictors has an R^2 of 0.5607.

From this it can be inferred that in the context of other predictors the predictor of the average sentiment of words in live chat message with a POS tag of other seems to explain more of the varience in the data than the average sentiment of words in live chat message.

Adding the final punctuation mark as a predictor and comparing it with the previous model in an ANOVA test results in a P value of $2.951 * 10^{-5}$. Since the p-value is less than the significance level, there is sufficient evidence to reject the null hypothesis, in favor of the alternative hypothesis. Therefore, it can be concluded that there is a statistically significant difference between the means of the populations. The model with final punctuation mark as a predictor has an R^2 of 0.64 while the previous model has an R^2 of 0.6108.

A. *Test Data Comparison*

As shown in table 2 all summary statistics are in favor of the new linear model in comparison to the VADER model. It

TABLE II
SUMMARY STATISTICS OF NEW LINEAR MODEL & NEW LEXICON BASED ON TEST DATA

Model	**MAPE**	R^2
New Linear Model	47.6296	0.6391
VADER model	73.7622	0.1908

can then be concluded that the linear model proposed in this paper is better for sentiment analysis of YouTube Live chat feed than the VADER model.

IV. DISCUSSION

The model discussed in this paper should, with the proper knowledge and tools, be able to be reproduced for any domain specific syntax in as short as eight hours. With current automated POS tagging and the lack of existing domain specific syntactical lexicons manual data annotation is needed for most domain specific syntax. This gives a distinct advantages to the PN-polarity models as they require less effort per data point. While there will always be a place for the plug and play aspects of VADER with a pre-built lexicon, many public projects exist where the current accuracy of general use lexicon models with pre-built lexicons is not appropriate for their use due to domain specific syntax, yet the general use lexicon models with the pre-built lexicon are still used.

A. *Limitations*

The major limitations to this paper's model are the small training set, single person bias for all data annotation. The training set used was only 1500 YouTube live chat messages. The lexicon's sentiment score and syntax could be greatly improved with more data. It must also be noted that many aspects of the model discussed in this paper are biased by the author. As there was only a single person was tagging both POS and sentiment. This makes it so that the model may be estimated not the true sentiment, but the person's perceived sentiment. While the true sentiment and the person's perceived sentiment may be highly correlated they are not the same. This could be solved with a larger and diverse team so that the perceived sentiment may average to the true sentiment.

V. CONCLUSION

This paper describes a multiple linear regression model and a new lexicon that can out preform the VADER model with a premade general lexicon for YouTube live chat sentiment analysis. Issues are apparent in the way that many use VADER with a premade general lexicon for domain specific syntactic environments. As shown in this YouTube case study the sentiment prediction in these domain specific syntactic environments can be dramatically improved with even a simple model and lexicons purpose built for them.

VI. References

References

[1] D. Santitto, “Twitch Chat Analyzer,” GitHub, Aug. 04, 2023. https://github.com/wredan/Twitch-Chat-Analyzer

[2] B. Thomas, “bthomas2622/twitchDisposition,” GitHub, Mar. 17, 2022. https://github.com/bthomas2622/twitchDisposition (accessed Mar. 05, 2024).

[3] opalsAndStones, “Made another WSB sentiment analyzing bot (I used machine learning) – 33

[4] T. Ray, “Using sentiment analysis to craft a narrative of the COVID-19 pandemic from the perspective of social media,” Jun. 2021.

[5] S.-M. Kim and E. Hovy, “Determining the Sentiment of Opinions,” Aug. 2004.

[6] M. Hu and B. Liu, “Mining and summarizing customer reviews,” Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’04, pp. 168–177, 2004, doi: https://doi.org/10.1145/1014052.1014073.

[7] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, “Building a Sentiment Summarizer for Local Service Reviews,” 2008.

[8] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.,” in LREC, 2006, pp. 417–422.

[9] C. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, no. 1, pp. 216–225, May 2014, doi: https://doi.org/10.1609/icwsm.v8i1.14550.

[10] A. Oad, “VADER Sentiment Analysis without and with English Punctuation Marks,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 2, pp. 1483–1488, Apr. 2021, doi: https://doi.org/10.30534/ijatcse/2021/1371022021.

[11] F. Singh, “Sentiment Analysis Made Easy Using VADER,” Analytics India Magazine, Dec. 08, 2020. https://analyticsindiamag.com/sentiment-analysis-made-easy-using-vader (accessed Mar. 05, 2024).

[12] J. Z. Maitama, N. Idris, A. Abdi, L. Shuib, and R. Fauzi, “A Systematic Review on Implicit and Explicit Aspect Extraction in Sentiment Analysis,” IEEE Access, vol. 8, pp. 194166–194191, 2020, doi: https://doi.org/10.1109/access.2020.3031217.

[13] ohn0, “ohn0/youtube-livechat-scraper,” GitHub, Dec. 19, 2023. https://github.com/ohn0/youtube-livechat-scraper (accessed Mar. 06, 2024).