Sentiment Analysis: Analyzing the sentiments of traffic related tweets.

Jonathan Suleman Ismail
MacEwan University

ABSTRACT

We present an informative method of analyzing traffic related tweets for the city of Edmonton while selecting upon three widely used tools with excellent F1 accuracy scores to demonstrate which of the tools selected can perform the best under the SEMEVAL metric for determining the average recall. We highlight possible challenges that can cause difficulties for measuring a sentiment analysis tools that focuses more on topicrelated tweets. We also provide possible solutions to increase effectiveness of the measured tools that produce a more accurate sentiment polarity classification.

1. INTRODUCTION

As the world population continues to grow and as society continues to expand and progress technologically, those involved in society need to adapt to these societal changes. It is the year 2018 and with its ever growing popularity, social media plat- forms have become the number one way of communication for most people around the world. This method of communication allows for people to stay in touch with family, engage in politi- cal discourse, provide an inside perspective of a faraway place or simply document their lives that highlight important milestones. Communication is now easily accessible by all and thus can have important impact on shaping society. Many business, social institutions and government bodies have utilized this form of communication to reach their mass audience. Businesses have used social media platforms to promote their products and services that naturally increase revenue and expand their business globally. According to research, businesses of the top 100 most valuable global brands that have utilized social media activity saw an 18% increase in their revenue for the previous year, while the least active, experienced a 6% decrease in revenue during the same period [3]. Social institutions regarding the social sciences have used social media to acquire information and conduct research to further investigate topics in their perspective fields. Sociologist and psychologist analyze user sentiments and study communication on social media to better predict and illustrate patterns of societal behavior [7]. What about government bodies?

In this study, I will be discussing a method of how government bodies can utilize social media to acquire information about the public on certain topics that can help them produce better policy that benefits society as a whole. The topic we will be discussing specifically is related to traffic in the city of Edmonton and the platform we will be using for the analysis is Twitter. The city of Edmonton can use this idea of observing social media discussion for providing solutions to straining problems, for example. . . creating more bike lanes, enforcing speed limits during populated events or prioritizing snow cleaning on certain streets to decrease accidents thus minimizing traffic. This information provides useful feedback for premature policy or can provide information for modifying existing ones. This feedback can also provide the city with what implementations are being perceived well and which implementations cause problems. The first goal for this study, and one of the most effective and important approaches to understanding the opinions of the public regarding social media is the use of sentiment analysis. The second goal for this study is to determine which of the many available sentiment analysis tools provide the highest level of accuracy in determining the sentiments of Edmontonians regarding traffic related tweets.

2. BACKGROUND AND RELATED WORK

Sentiment analysis is a technique that uses natural language processing, statistics, and/or machine learning methods to extract,

identify, and characterize sentiment in a specific text. These characteristics include a mixture of feelings, attitudes, emotions, and/or opinions. There are many Sentiment Analysis techniques that have been used by business establishment, social institutions and government bodies that help them better understand the sentiments of their audience and provide an appropriate solution in response to the data collected.

3. METHODS

3.1 Objective

Social networking and microblogging sites such as Twitter provide rich and complex data for researchers that can be assessed in a comparable informative way. When companies, researchers or in our case, municipal bodies want to understand the general public on a social media platform, they do so by means of sentiment analysis. Our objective in this study is to determine which of the tools examined will provide the highest level of accuracy in determining the sentiments of Edmontonian regarding traffic related text on Twitters' API. Understandably, the effects and accuracy of sentiment analysis relies heavily on the context or methods of which it is being conducted. We will provide a series of steps and create a method that will help us better understand the tools in determining a solution for our thesis. Before we begin, we will need a set of tools that can help us in our sentiment analysis. We will begin by describing sentiment analysis and present and overlay of the types of techniques available for performing the analysis. Sentiment analysis is a process of computationally identifying expressive sentiments in a given text. Many tools have been developed over the years for analyzing sentiments in short informal social media texts with various techniques used to preform the analysis. When performing sentiment analysis, there are three techniques used by researchers, a lexicon approach and a machine learning approach and a hybrid approach[5]. The lexicon based analysis is a collection of precompiled sentiments which use semantic or statistical methods to find sentiment polarity of a given text . Whereas, a machine learning approach focuses on creating a model by training classifiers with labeled example. This method will require examples of positive, negative and neutral classes which involve training the algorithm based on a given dataset. According to research, this method can be troublesome for a few reasons; firstly, it requires extensive data training and secondly, it derives features "behind the scenes" [5] which make modifications and generalization difficult. The hybrid approach encompasses the combination of both lexicon and machine learning techniques.

3.1.1 Potential Error We will highlight common problematic challenges encountered by researchers when evaluating sentiment analysis tools. Tweets containing mixed sentiments which incorporate both positive and negative sentiments within 140 characters can be considered difficult to analyze due to the limitations of articulating complex opinions. [1]. These erroneous tweets can be caused by many factors such as mixed emotions, neutral-sentiment questions or sentences that contained a rhetorical, sarcastic or satirical structure [1]. Therefore, in order to obtain a high performing accurate tool, we will need to overcome these intricate erroneous features.

3.2 Sentiment Analysis Tools

Researchers have developed a high performing tool called VADER which was compared against seven well established sentiment analysis tools [6]. VADER includes many lexical features such as Western-style emoticons, sentiment-related acronyms and commonly used slang. The premise of the study was comprised of constructing a valence-aware sentiment lexicon, which established context awareness. The tools examined by the researchers not only determine polarity but also included the strength of the sentiment expressed. As a result, researchers examined that VADER greatly outperformed its competitors, including individual human raters. The runner-up Hu-Liu04, scored relatively high as shown in Table 1, however, in comparison to VADER scored much lower. The low score was caused by incorrect classification, presumably due to lack of coverage for social media orientated language that include emoticons, slang and abbreviated texts.

We will examine another study that conducts a thorough comparison of sentiment analysis tools and provides a comprehensive benchmark. Several tools were evaluated by researchers that examined the strengths and weaknesses of methods used to determine the viability of a sentiment analysis. The methods presented by researchers compared twenty-four tools across eighteen datasets – eight of which are twitter based. The essential function across all measured tools was having the ability to perform sentence-level analysis which provides polarity detection in short messages that are commonly used in social media.

The goal of this study was to create a benchmark that provides a descriptive comparison of widely used sentiment analysis tools. The researchers compared the tools in two separate experimental testbeds. The first was set to identify the performance of a 2-class(positive/negative) function and the second was set to identify the performance of a 3-class(positive/negative/neutral) function. For our case, we will focus on the 3-class sentiment results. The study suggests that the quality of Twitter sentiment analysis tools varies considerably, and that the performance variation can have important implications for various Twitter-based analytics. Based on the results of the study, the best overall 3-class performing tools are VADER, LIWC15 and AFINN which performed exceptionally on all datasets including the differing twitter datasets shown in Figure 1 and Table 3.

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)				
Ind. Humans	**0.888**	0.95	0.76	0.84
VADER	0.881	**0.99**	**0.94**	**0.96**
Hu-Liu04	0.756	0.94	0.66	0.77
SCN	0.568	0.81	0.75	0.75
GI	0.580	0.84	0.58	0.69
SWN	0.488	0.75	0.62	0.67
LIWC	0.622	0.94	0.48	0.63
ANEW	0.492	0.83	0.48	0.60
WSD	0.438	0.70	0.49	0.56

Figure 1. VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines

3-Classes		
Pos	**Method**	**Mean Rank**
1	VADER	4.00(4.17)
2	LIWC15	4.62
3	AFINN	4.69
4	Opinion Lexicon	5.00
5	Semantria	5.31
6	Umigon	5.77
7	SO-CAL	7.23
8	Pattern.en	9.92
9	Sentiment140	10.92
10	Emolex	11.38
11	Opinion Finder	13.08
12	SentiWordNet	13.38
13	Sentiment140_L	13.54
14	SenticNet	13.62
15	SentiStrength	13.69(13.71)
16	SASA	14.77
17	Stanford DM	15.85
18	USent	15.92
19	NRC Hashtag	16.31
20	LIWC	16.46
21	ANEW_SUB	18.54
22	Emoticons	21.00
23	PANAS-t	21.77
24	Emoticons DS	23.23

Figure 2. The mean rank table for all datasets.

The tools that we will select for our research upon examination of the previous studies will be VADER, LIWC15 and AFINN. All the mentioned tools have performed remarkably overall however our focus was on the twitter datasets. Our aim was to select tools that can perform exceptionally under 140 characters of short condensed texts while simultaneously articulating a wide range of sentiments. All three tools have been ranked as the top of social networking datasets. The reasons for not including Umigon as a tool of interest, is due to its low lexicon size which can be problematic for analyzing specifically detailed tweets [1]. The tools selected also present promising features such as emoticon and slang analysis [6]. The tools would also be ideal because they contain a large lexicon dictionary size and provide great analysis for 3-class functions and therefore would be great candidates for operating sentiment analysis on traffic related tweets.

3.3 Dataset and Methods

Beyond the studies examination, we will comprise a series of methods to determine which upon the three selected tools will outperform its competitors for our specified analysis. In order to perform a test between the tools, we will need to filter key points of interests in our test bed. The target of this study is to establish a tool by analyzing traffic related tweets in Edmonton. While many studies focus on collecting large amounts of data and generalize their findings, we will establish a more specific dataset by collecting specific topic related posts. In order to do

3-Classes		
Pos	**Method**	**Mean Rank**
1	Umigon	2.57
2	LIWC15	3.29
3	VADER	4.57(4.57)
4	AFINN	5.00
5	Opinion Lexicon	5.57
6	Semantria	6.00
7	Sentiment140	7.00
8	Pattern.en	7.57
9	SO-CAL	9.00
10	Emolex	12.29
11	SentiStrength	12.43(11.60)
12	Opinion Finder	13.00
13	SentiWordNet	13.57
14	SenticNet	14.14
15	SASA	14.86
16	LIWC	15.43
17	Sentiment140_L	15.43
18	USent	16.00
19	ANEW_SUB	19.14
20	Emoticons	19.14
21	Stanford DM	19.43
22	NRC Hashtag	20.00
23	PANAS-t	20.86
24	Emoticons DS	23.71

Figure 3. The mean rank table for datasets of Social Networks.

so, preprocessing the data would be required. Researchers suggest adopting a bounding box approach for filtering the dataset [4], thereby tweets will be collected from a specific geographic regions; in our case the city of Edmonton. Once a location has been established we will need to preprocess the data fur- ther by removing all non English words from the database. Lisa Branz in her study [4] of filtering social network data has pro- posed a hypothesis, that precise and faceted classification was not necessary. She suggested a simple method in which all re- lated subjects would be classified into a general word. In her example, she differentiated the topic of sports and non-related sports by assigning tweets to either "Sports" or "Other". This type of classification greatly narrowed her dataset and only in- cluded sport related tweets. To replicate her tapered dataset, we will use this method and specifying our tweets by classifying all related traffic tweets to "Traffic" and assign non-traffic related tweets to "Other". By utilizing this simple classification method and applying it to the geographic filtered region, our dataset will contain English traffic-related tweets from Edmonton. When researchers evaluated the performance of the VADER tool, [5] they found that 400 tweets was sufficient in determining the accuracy of the tool. We will adopt this figure in determining which of our tools selected, can perform sentiment analysis on traffic related tweets accurately.

3.4 Procedure and Performance Measures

Now that we've selected our tools, we will proceed in collecting our dataset from the Twitter API. We will use the filtering method mentioned above to gather traffic related tweets in Edmonton. Upon gathering our dataset, we will use the given procedures to determine which of the tool selected will outperforms its competitors in terms of accuracy for traffic related tweets. According to a study [2], a key aspect in evaluating sentiment analysis methods consists of using accurate gold standard labeled datasets. These gold standards would require human labeling which would then be later compared with the analysis of a tool. However, some errors and concerns are in place. According to the study [2], human labelling is considered to be subjective, therefore the researchers implemented an agreement strategy in which evaluators would vote in favor of the majority. This voting would then ensure that each tweet or sentence had an agreed-upon polarity assigned to it. We will begin our procedure by applying sentiment analysis to the given dataset by each tool. The comparisons between the tools and the gold standard labels would quantify an accuracy estimation. Comparing each tweet with the gold standard label would identify which tool correctly identified the sentiments based on the gold standard labelling. The measurement used to quantify the accuracy is derived by the International Workshop on Semantic Evaluation – 2017 (SemEval): Subtask A, which given a tweet, decides whether it expresses positive, negative or neutral sentiment. The formula below will calculate average recall across the positive, negative and neutral classes. The average ranges from [0,1], where the value of 1 is achieved by a perfect classifier.

$$AvgRec = \frac{1}{3}(R^P + R^N + R^U) \quad (1)$$

Each tool will be assessed based on the given comparative metrics and will be evaluated based on their Average Recall (Av-

gRec) score. The overall accuracy would be determined by calculating the Average Recall (AvgRec) score between the measured tools and the gold standard labels. The accuracy mean of each tool would then be compared to each other, and the tool whose performance outdistanced its competitors would be ranked as the better tool.

4. RESULTS

The Recollection metric for positive, negative and neutral classes are based on the amount of correct sentiments compared to the gold standard label. Recall is the number of true classifications divided by the total number of elements that are known to belong to a specific class; thus a low recall indicates that some known elements of a class are missed.

	VADER	LIWC	AFINN	Gold Label
Positive	20	25	24	35
Negative	195	141	92	262
Neutral	31	37	54	108
Ranking:	1	1	2	–

Table 1*: Each tool has assigned Positive, Negative and Neutral sentiments to the Twitter dataset. The Sentiment scores of the evaluated tools are compared to the Gold Standard Labels which are the correct assigned scores.*

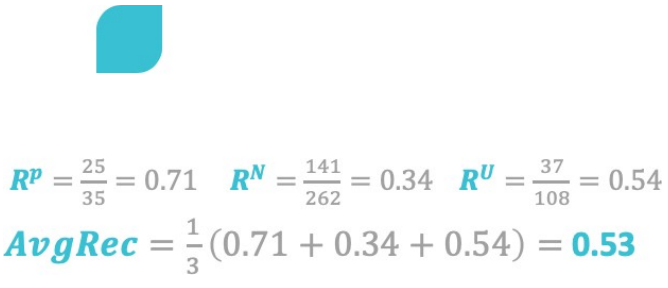

Figure 4. LIWC

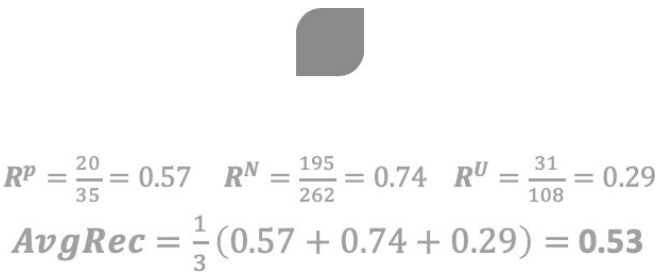

Figure 5. VADER

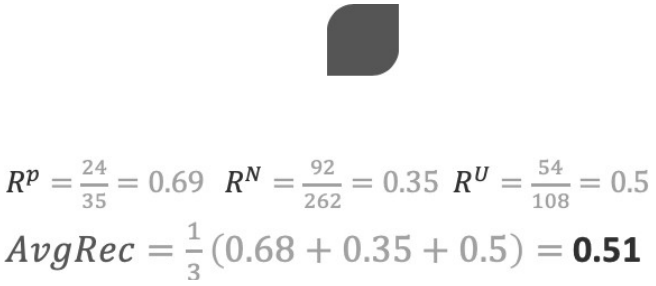

Figure 6. AFINN

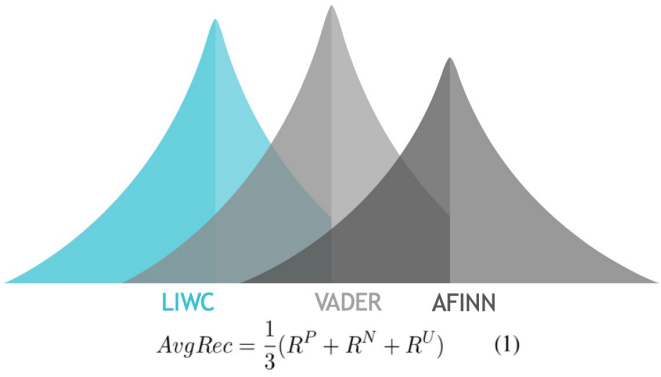

Figure 7. The formula above is the measurement used to quantify the accuracy that is derived by the International Workshop on Semantic Evaluation – Subtask A and the graph demonstrates the ranking of the tools.

The formula developed by SemEval calculates the average recall of a class and determines the percentage of accurate classes correctly classified[6]. The formula consists of P, N and U which refer to recall with respect to positive, negative and neutral classes respectively. Table 1 and Figure 7 illustrates the average recall of VADER, AFFINN and LIWC15 and demonstrates that VADER and LIWC had a average recall of 53% while AFFINN had an average recall of 51%; the higher the better. As you can see, the tools are placed in the same ranking as the previous studies we've examined, placing VADER at the top. However the recall accuracy is just over 50 percent which is not great, meaning almost half of the tweets evaluated incorrectly categorized the sentiments based on the gold standard labels given. The top ten tools that participated in SemEval-2017 averaged over 64% which in comparison to our results, yields a 10% difference in recall accuracy.

5. DISCUSSION

The effects and the accuracy of sentiment analysis relies heavily on the context of which it's being used. Although the tools evaluated by past researchers produced excellent results, they were not faced with similar challenges. The biggest challenge faced, were the weak gold standard labels. The researchers who developed VADER's sentiment analysis tool used crowd sourcing marketplace websites like Amazon Mechanical Turk which used 20 human raters to rate the sentiment polarity of their twitter dataset. Our study used 3 human raters which greatly affected the results. For example, several tweets had three different sentiment polarity assigned to them by three different raters. This uncertainty was verified by a fourth party to determine the favored sentiment polarity. If many raters were to evaluate these tweets, our gold standard labels would be more extensive thus producing a more accurate result. Another challenge was determining the context of the tweets that were being evaluated. Determining the sentiment polarity of the tweet seemed to be quite problematic. For example, tweets that contained sarcasm and/or mixed emotion were hard to categorize for both humans and tools. Specific traffic related words such as "accident" caused confusion between the raters and the individual tools as well, therefore establishing a framework of predefined words with attached polarity scores would increase the accuracy of the tools evaluated. Due to these challenges, the overall performance may not accurately represent the accuracy of the tools evaluated. But however has highlighted some key aspects to consider for future work in determining a tool for analyzing traffic related tweets in Edmonton. Luckily, all three tools have modifiable lexicon dictionaries and can be tailored to work with more specific topics.

6. CONCLUSION

The study illustrates the ranking and performance between three widely respected tools and provides a general review of the research process while summarizing the methods used in the determining a prospective sentiment analysis tool. In essence this paper illustrates the complexity of sentiment analysis and highlights the factors and problematic hurdles that can occur while measuring the accuracy of a tool. The results, as expected where consistent with past research in terms of ranking and averaged slightly over 50% recall accuracy. The tools when compared to the SEMEVAL results performed well considering the unmodified dictionaries and the specificity of traffic related tweets. We hope our study not only helps researchers but also government bodies select or develop a more topic-specific sentiment analysis tool that can assist in achieving future endeavors.

7. REFERENCE

[1.] Abbasi,, Ahmed, et al. "Benchmarking Twitter Sentiment Analysis Tools."

[2.] Ribeiro, Filipe N, et al. "SentiBench - a Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods." 14 July 2016.

[3.] Chen, Hsuanwei Michelle, and Patricia C Franks. "Exploring Government Uses of Social Media through Twitter Sentiment Analysis." Journal of Digital Information Management, vol. 14, no. 5, 4 June 2016.

[4.] Branz, Lisa, and Patricia Brockmann. "Poster: Sentiment Analysis of Twitter Data: Towards Filtering, Analyzing and Interpreting Social Network Data." The 12th ACM International Conference on Distributed and Event-Based Systems, doi.org/10.1145/3210284.3219769

[5.] Pawar, K. K., Shrishrimal, P. P., & Deshmukh, R. R. (april 2015). Twitter Sentiment Analysis: A Review. International Journal of Scientific & Engineering Research, 6(4).

[6] Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. Association for Computational Linguistics, 502-518.

[7] Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. Big Data & Society, 1(15).