INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING & COMPUTER SYSTEMS (IJSECS)
ISSN: 2289-8522 e-ISSN: 2180-0650
VOL. 9, ISSUE 1, 1 – 9
DOI: https://doi.org/10.15282/ijsecs.9.1.2023.1.0105

ORIGINAL ARTICLE

A Domain-Specific Evaluation of the Performance of Selected Web-based Sentiment Analysis Platforms

Manuel O. Diaz Jr.[1]

[1]West Philippine Sea Institute, Apple One Equicom Tower, Biliran St., Cebu City, Philippines 6000.

ABSTRACT – There is now an increasing number of sentiment analysis software-as-a-service (SA-SaaS) offerings in the market. Approaches to sentiment analysis and their implementation as SA-SaaS vary, and there really is no sure way of knowing what SA-SaaS uses which approach. For potential users, SA-SaaS products are black boxes. Black boxes, however, can be evaluated using a set of standard input and a comparison of the output. Using a test data set drawn from human annotated samples in existing studies covering sentiment polarity of news headlines, this study compares the performance of selected popular and free (or at least free-to-try) SA-SaaS in terms of the accuracy, precision, recall and specificity of the sentiment classification using the black box testing methodology. SentiStrength, developed at the University of Wolverhampton in the UK, emerged as consistent performer across all metrics.

ARTICLE HISTORY
Received: 10 March 2022
Revised: 11 October 2022
Accepted: 27 October 2022
Published: 1 January 2023

KEYWORDS
Sentiment analysis
Sentiment polarity
Polarity identification
Analysis-as-a-service
Text classification

INTRODUCTION

We humans acquire language skills growing up. We learn them informally at home and through our various interactions, and we learn them formally in school. We learn to write in and speak a language. Language is a structured system of communication, and when it has evolved and evolves over time through repetitive or continuous use, it is referred to as natural language. The use of computers to process natural language data is called natural language processing (NLP), the ultimate goal of which is that we can communicate to computers using natural language and computers will be able to understand us. NLP is therefore a computer science problematization of the various aspects of linguistics, such as phonetics, phonology, morphology, syntax, semantics, and pragmatics.

A subset of NLP is sentiment analysis. Sentiment can be characterized as positive or negative evaluation of something expressed through language [1] and sentiment analysis is concerned with how sentiment is expressed specifically in written text.

Text contains information which can be either objective or subjective. Objective information is factual and does not express sentiment. On the other hand, subjective information may express personal feelings, views, opinions or beliefs [2]. When a statement expresses personal feelings, views, opinions or beliefs, the statement expresses an evaluation of the subject, which could be positive or negative. The positiveness or negativeness of a statement relative to a subject is referred to as sentiment polarity. It can be viewed as dichotomous: a statement is either positive or negative. It is exclusive and is either one or the other. However, it can also be viewed as being a range: a statement can be very positive, while another can just be more positive than negative, but close to being neutral.

Sentiment polarity identification is just one of the many tasks of sentiment analysis. According to Taboada [1], approaches and terminologies may vary, but the main goal of sentiment analysis is to determine whether a text, or a part of it, is subjective or not and, if subjective, whether it expresses a positive or a negative view. Esuli and Sebastiani [3] defines this as having three different aspects, which are subjectivity, polarity and strength.

Subjectivity of text refers to whether text expresses facts about entities, events and their properties or it describes people’s sentiments, appraisals or feelings toward entities, events and their properties [2]. As a text classification problem, [4] it is more difficult than distinguishing between positive, negative and neutral sentiments.

Polarity of text is its orientation towards being positive or being negative, also known as semantic orientation. When a word is often used to convey favorable sentiment or evaluation of a subject, its semantic orientation is said to be positive. Examples are diligent, wise, happy. On the other hand, if a word is often used to convey unfavorable sentiment or evaluation of a subject, its semantic orientation is said to be negative. Examples are boring, arrogant, cruel. Semantic orientation is generally conveyed using adjectives [1], but other parts of speech such as noun, verb, adverb and phrases that contain them can also effectively convey the same [5].

The strength of opinion in text refers to the degree of polarity of the text. Two statements can express a positive sentiment, but one can be more positive than the other. Consider for example a Facebook post which has received multiple Likes. The Likes indicate positive opinion on the post, but reading through the comments of those that Liked the post can show different degrees of Liking. Similarly, consider product reviews where customers rate products using n-star ratings where 1 star is the lowest rating and 5 stars is the highest. Technically, a 1-star rating is still positive, but again it is not

***CORRESPONDING AUTHOR** | M. Diaz | ✉ mdiaz@westphilippinesea.org

unusual to expect that a review that accompanies a 1-star rating will not be as flattering as a review that accompanies a 4-star rating. One approach to this problem is by estimating the sentiment strength of user reviews according to the strength of adverbs and adjectives expressed by users in their opinion phrases [6].

The identification of sentiment polarity finds application in business strategy, where business organisations are given the opportunity to gauge how their products are received in the market. Traditionally, sales is a very good indicator of market response, e.g. if it is selling then it must be good, but understanding market sentiment provides businesses with better insight so they can better position their products and services. As such, some of the most studied texts in sentiment analysis are product and movie reviews [4]. Politics, where public opinion matters, is also a good ground for sentiment analysis. Researchers have also conducted an analysis of political sentiment orientations on Twitter relative to the General Elections of India in 2019 [7], while sentiment analysis has been used to determine whether individual conversational turns in U.S. Congressional floor debates support or oppose some given legislation [8]. Where opinion is expressed in written text, sentiment analysis will most likely find application.

Meanwhile, interest in sentiment analysis has steadily increased over time in and outside of academia. Leveraging the new business models of cloud services, sentiment analysis has become a viable commercial offering as-a-service and an important tool for competitive advantage. Much of the literature on sentiment analysis, however, looks at the various approaches to doing sentiment analysis and how a chosen approach is better suited to a task, how it performs, or how it can be improved. It has articulated a myriad of ways by which algorithms can exert power through decisions they make in the prioritization, classification, association, and filtration of information [9]. Algorithms behind sentiment analysis platforms exert the same power, particularly if the sentiment classification results in decisions being made. With as-a-service platforms, however, one cannot look under the hood, so to speak. Like black boxes, we can only observe their output for a given set of inputs, or we simply rely on the claims by the party marketing it to the public. The best way to get a feel of their performance is, of course, to try them out. In this study, this limitation imposed by as-a-service platforms requires the use of black-box testing, a software testing strategy looks at what inputs are available and what output is expected given those inputs.

As-a-service platforms make sentiment analysis available to a very wide audience. These then need to cater to multiple domains. They are sometimes called general purpose sentiment analyzers, capable of doing cross-domain sentiment analysis. Cross-domain sentiment analysis is the application of sentiment analysis to more than one domain. These domains may be very different from each other. Some domains are more studied than others, so there are more training datasets about those domains that algorithms can learn from. On the other hand, datasets on other domains can be sparse, if not available at all. Thus, cross-domain sentiment analysis continues to be a challenge. A sentiment analyzer is usually optimized for a particular domain. One sentiment analyzer may be good within a domain but may be wanting in another domain. In this study, the sentiment analyzers will be tasked with the analysis of news headlines. Headlines do not convey opinion, but can nonetheless be replete with sentiment, but one may have to read between the lines. As a sentiment analysis task, it is not just domain-specific. The very nature of news headlines adds a certain degree of complexity.

METHODOLOGY

In brief, the methodology used in this project consists of the following: first, the identification of the web-based software-as-a-service (SaaS) to consider; second, preparation of the test data to use; third, establishment of the evaluation metrics; fourth, the conduct of the actual test; and finally, discussion of the results.

Identification of Sentiment Analysis SaaS (SA-SaaS)

There are only a handful of studies focused on SA-SaaS. One analyzed eight online programs for sentiment analysis that can be accessed for free and which can, on the basis of their algorithm, give a positive, negative or neutral opinion of a text [10]. Four were API-based, such as Sentigem: SentimentAnalysis API, Text Analytics & Sentiment Analysis API, Google Cloud Natural Language API, Microsoft TextAnalytics API, while the other four provided a web user interface such as Daniel Soper Sentiment Analyser, SentiStrength, IntentCheck and Sentiment Tool. A more comprehensive comparative analysis of 15 web services, namely Alchemy API, Lymbix, Musicmetric, Openamplify, Opinion Crawl, Opendover, Repustate, Semantria, Sentiment140, Sentiment Analyzer, SentiRate, Sentimetrix, Uclassify, ViralHeat, and Wingify has also been attempted [11]. Following these works, this study also aims to evaluate commercially available as-a-service offerings, but distinguishes itself from the aforementioned studies based on two primary criteria for sampling:

1. The SA-SaaS includes a web user interface (UI), i.e. it can be used by anyone with access to the Internet using a browser without having to code as in the case of web services that are only available using an applications programming interface or API; and
2. The SA-SaaS is free to use, or if it is a paid service, it has a demo which anyone with access to the Internet using a browser can try out.

Commercially available web services generally differentiate themselves from the competition with proprietary algorithms. Each of these service providers will claim advantage over the other, but the only way to find out which platform provides the service that will address particular requirements and deliver best results is to compare them using the same baseline data set, which is exactly what this study will demonstrate.

Using the twin criteria of availability of web UI and that the service is free at least to try, Daniel Soper Sentiment Analyzer and SentiStrength are good candidates [10]. The others are API-based, and those that are not could not be accessed as of 16 September 2021. In the web services another group of researchers looked at, only Sentiment Analyzer and Uclassify satisfy the twin requirement [11].

Two more platforms were selected using a Google search. The results from the Google search included links to actual sentiment analysis platforms and blogs listing supposed top sentiment analysis platforms. These were checked against the twin criteria of availability of web UI and that the service is free or at least has a free demo. Text2Data and Komprehend.io were then selected for this study.

Daniel Soper Sentiment Analyzer

Created by Professor Daniel Soper, Vice-Chair, Department of Information Systems & Decision Sciences in the College of Business and Economics at the California State University, Fullerton, Sentiment Analyzer is a free tool for conducting sentiment analysis on virtually any text written in English. It computes a sentiment score which reflects the overall sentiment, tone, or emotional feeling of the input text. Designed to be a general-purpose sentiment analysis tool, it is not oriented toward any specific domain. It has been trained using the collection of more than 8,000 writing samples and transcripts of spoken conversations from the American National Corpus (ANC) which contains writing samples from a wide variety of genres and domains.

SentiStrength

SentiStrength was developed by Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Di Cai at the Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton in the UK as part of the Cyber Emotions Project funded by the European Union. It uses a dictionary of sentiment words with associated strength measures and exploits a range of recognised non-standard spellings and other common textual methods of expressing sentiment. Developed through an initial set of 2,600 human-classified MySpace comments, it has been evaluated on a further random sample of 1,041 MySpace comments [12].

uClassify

Developed by a small Stockholm-based team led by Jon Kågström, uClassify has in its core a multinomial Naive Bayesian classifier. Kågström and team claims to have added a couple of steps that further improves the classification, which results in probabilities from 0 to 1 that a document belongs to a class, enabling setting of a threshold for classifications.

Text2Data

Text2Data is a text analytics SaaS startup based in London, UK. The platform uses NLP and Deep Learning along with proprietary algorithms, and is fully cloud-based. To analyze text, it splits every sentence into chunks and represents them as a tree structure. Probabilistic methods based on pre-trained data models are then used to get the final score.

Komprehend.io

Komprehend.io is an Indian startup founded by Angam Parashar, Ankit Narayan Singh and Muktabh Mayank. It uses Long Short Term Memory (LSTM) algorithms to classify text, modelling sentences as a chain of forget-remember context-based decisions. Komprehend.io has been trained on social media and news data and various custom datasets for different clients, including handling of casual and formal language.

Preparation of the Test Data

Meanwhile, the test data is built based on sample annotated data from existing works [13], [14] and [15] which looked at the sentiment polarity of news headlines. Headlines are particularly interesting because they are usually limited in length to a sentence at most. Headlines convey a lot using as few words as possible, while at the same time being interesting enough to catch the attention of readers. The test data is listed below, annotated with the study from which the headline was sourced together with the expected sentiment polarity that the study it was sourced from tagged the particular headline.

Table 1 lists news headlines that have been manually annotated as text with positive sentiment polarity as used in published studies. Likewise, Table 2 lists news headlines from the same studies, but were manually annotated as text with negative sentiment polarity.

Table 1. Sample news headlines with positive polarity from published studies.

ID	Headline Text with Positive Polarity
1	CUNY Professor Criticizes Jews [15]
2	To Allay Fears of Islam, Mosques Invite Visitors [15]
3	Police investigate fire at Islamic community centre in Muswell Hill [15]

4	Celebrating 50 Issues of Jewish Socialist Magazines [15]
5	Ta da! New Guinness World Record set for completing a Rubik’s Cube in just 3.253 seconds... by a robot [14]
6	J.K Rowling delights Harry Potter fans by posting 2,400 word History Of The Quid-ditch World Cup on Pottermore [14]
7	He’s a real-life hero! Liam Neeson saves stray dog from ‘teenagers throwing stones at it’ [14]
8	Caravan of love: Britain’s youngest parents enjoyed romantic beach break together just a month before their baby was born and gave each other “True Love” bracelets [14]
9	Radical policies to end the failure of community care [13]
10	Measles jab suspected in autism cases [13]
11	BSE epidemic ‘could be over by mid-1998’ [13]
12	Polio and measles ‘to be erased by 2010’ [13]
13	Treatment offers hope for ‘human vegetables’ [13]

It should be noted that if humans are asked to manually annotate statements with their sentiment polarity, there will be some level of disagreement. According to Plank, et. al. (2014), human annotators agree up to 80% of the time. Applied in the current project, it can then be inferred that of the 26 test data items, even human annotators may disagree with up to 5 items or reach a consensus only for 21 of the 26 items.

Table 2. Sample news headlines with negative polarity from published studies.

ID	Headline Text with Negative Polarity
14	Muslim Gave Racist Speech, Jackson says [15]
15	Appeals Court Upholds Terrorist Label for a Jewish Group [15]
16	Suspected Islamic terrorists arrested [15]
17	Tensions as Jewish settlers press demands in Hebron/Israel and West Bank Politics [15]
18	US Army Fort Hood murder-suicide: Soldier kills three [14]
19	Meat from cattle slaughtered in ‘cruel’ kosher ceremony is in your high street burger [14]
20	The shock troops sent to terrorise Putin’s enemies as Crimea prepares to vote on join-ing Russia [14]
21	Room for a little one? SIX-seater buggy spotted being pushed around Cambridge confusing tourists and locals [14]
22	Psychiatric bed cuts ‘could lead to more murders’ [13]
23	Judge warns of mentally ill nightmare [13]
24	Victim's agony as senile sex beast is freed [13]
25	Health shake-up after ‘care in community’ fails [13]
26	Inquiry after mentally ill man was set free to kill [13]

A more robust evaluation of machine learning platforms such as those that do sentiment analysis and classification will generally require huge data test sets. However, these types of evaluations are generally conducted when the code and algorithms used are disclosed. In a more practical sense, such as when evaluating commercial off-the-shelf software, and in this case as-a-service products, only a select test data set is employed to determine whether a product really does what it claims it can do. In fact, using the black-box software test strategy, test analysts simply define test data based on scenarios where the product should succeed and scenarios where the product may fail. Products that fail on scenarios which they are expected to succeed fail the evaluation, while products that succeed on scenarios which they may fail pass the evaluation, and are even hailed as being robust, fail-safe or failure-resilient. The use of 26 data points in this study, sourced from samples in 3 prior studies conducted, provides sufficient variety of test data for purposes of black-box testing.

Establishment of Test Metrics

In this study, the test metrics used are accuracy, precision, recall or sensitivity, and specificity, which have now become standard measures of performance in classification tasks. Accuracy is the ability of a system to correctly classify sentiments. That is, what are supposed to be positive texts are classified as positive texts, and what are supposed to be negative texts are classified as negative texts. Accuracy is quantified by the ratio between the total number of correctly classified items and the total number of items classified:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

True Positive are items expected to be classified as positive and are correctly classified as positive, *False Positive* are items expected to be classified as negative but are incorrectly classified as positive, *True Negative* are items expected to be classified as negative and are correctly classified as negative, and *False Negative* are items expected to be classified as positive but are incorrectly classified as negative.

Precision quantifies the quality of performance in terms of correct positive classifications. It is indicative of the proportion of positive classifications that are actually correct. It is the ratio between the number of correct positive classifications and the total number of positive classifications made:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Meanwhile, quantifies the quality of performance in terms of actually correct positive classifications. It is a measure of how much of the items expected to be classified as positive were actually classified as positive. It is also known as sensitivity. It is the ratio between the number of correct positive classifications and the total number of expected positive classifications, and is mathematically expressed as:

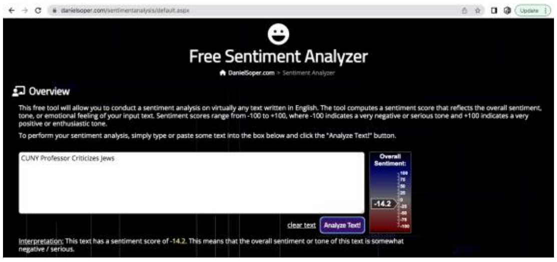
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Specificity quantifies the quality of performance in terms of actually correct negative classifications. It is a measure of how much of the items expected to be classified as negative were actually classified as negative. It is the ratio between the numer of correct negative classifications and the total number of expected negative classifications, and is mathematically expressed as:

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Conduct of the Actual Test

Each of the test data consisting of 26 news headlines listed in Table 1 and Table 2 were manually fed to the web-based sentiment analysis platforms. Shown in Figure 1 below is Daniel Soper showing the sentiment classification results for news headline #1, “CUNY Professor Criticizes Jews.” Figure 2 shows the sentiment classification results from SentiStrength on news headline #6, “J.K Rowling delights Harry Potter fans by posting 2,400 word History Of The Quid-ditch World Cup on Pottermore.” The Daniel Soper UI looks sleek and modern. The SentiStrengh UI, on the other hand, is quite basic reminiscent of pre-interactive web era. However, the results from the former is just straightforward sentiment score, while the latter provides an approximate rationale for the classification.



danielsoper.com/sentimentanalysis/default.aspx

Free Sentiment Analyzer

DanielSoper.com > Sentiment Analyzer

Overview

This free tool will allow you to conduct a sentiment analysis on virtually any text written in English. The tool computes a sentiment score that reflects the overall sentiment, tone, or emotional feeling of your input text. Sentiment scores range from -100 to +100, where -100 indicates a very negative or serious tone and +100 indicates a very positive or enthusiastic tone.

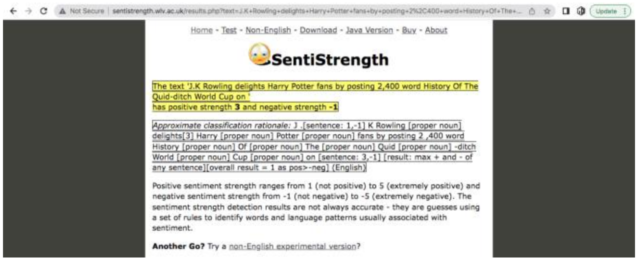
To perform your sentiment analysis, simply type or paste some text into the box below and click the "Analyze Text!" button.

CUNY Professor Criticizes Jews

Overall Sentiment: -14.2

Interpretation: This text has a sentiment score of -14.2. This means that the overall sentiment or tone of this text is somewhat negative / serious.

Figure 1. Daniel Soper



Not Secure | sentistrength.wlv.ac.uk/results.php?text=J.K+Rowling+delights+Harry+Potter+fans+by+posting+2%2C400+word+History+Of+The+...

SentiStrength

The text 'J.K Rowling delights Harry Potter fans by posting 2,400 word History Of The Quid-ditch World Cup on '
has positive strength **3** and negative strength **-1**

Approximate classification rationale: J .[sentence: 1,-1] K Rowling [proper noun] delights[3] Harry [proper noun] Potter [proper noun] fans by posting 2 ,400 word History [proper noun] Of [proper noun] The [proper noun] Quid [proper noun] -ditch World [proper noun] Cup [proper noun] on [sentence: 3,-1] [result: max + and - of any sentence][overall result = 1 as pos>-neg] (English)

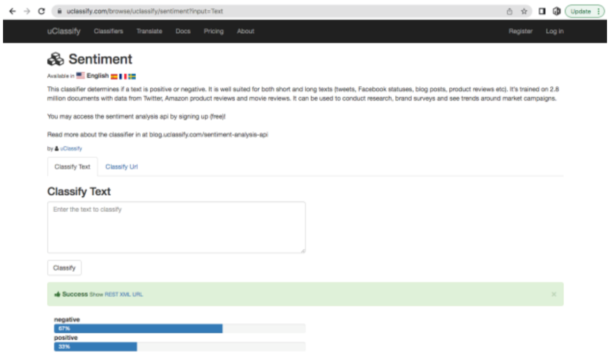
Positive sentiment strength ranges from 1 (not positive) to 5 (extremely positive) and negative sentiment strength from -1 (not negative) to -5 (extremely negative). The sentiment strength detection results are not always accurate - they are guesses using a set of rules to identify words and language patterns usually associated with sentiment.

Another Go? Try a non-English experimental version?

Update
Home - Test - Non-English - Download - Java Version - Buy - About

Figure 2. SentiStrength

Figure 3 below shows the results for uClassify, working on the classification of the 23rd news headline from the test set, “Judge warns of mentally ill nightmare.” Instead of polarity scores, it provides percentages which are more intuitive than using a range of classification scores. Meanwhile, Figure 4 shows the results for Text2Data, fed with the 10th news headline from the test set, “Measles jab suspected in autism cases.”



uclassify.com/browse/uclassify/sentiment?input=Text

Sentiment

Available in English

This classifier determines if a text is positive or negative. It is well suited for both short and long texts (tweets, Facebook statuses, blog posts, product reviews etc). It's trained on 2.8 million documents with data from Twitter, Amazon product reviews and movie reviews. It can be used to conduct research, brand surveys and see trends around market campaigns.

You may access the sentiment analysis api by signing up (free)!

Read more about the classifier in at blog.uclassify.com/sentiment-analysis-api

by uClassify

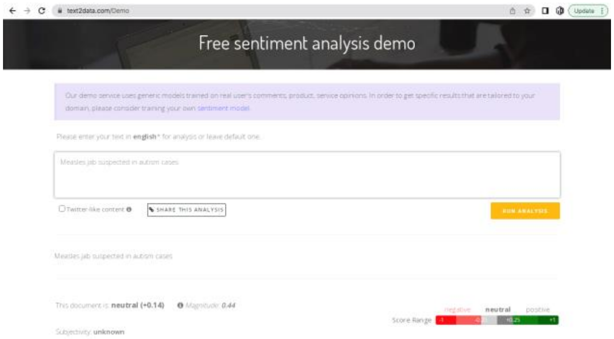
Classify Text

Enter the text to classify

Success Show REST XML URL

negative 67%

positive 33%



text2data.com/Demo

Free sentiment analysis demo

Our demo service uses generic models trained on real user's comments, product, service opinions. In order to get specific results that are tailored to your domain, please consider training your own sentiment model.

Please enter your text in **english*** for analysis or leave default one.

Measles jab suspected in autism cases

□ Twitter-like content

Measles jab suspected in autism cases

This document is: **neutral (+0.14)**

Magnitude: 0.44

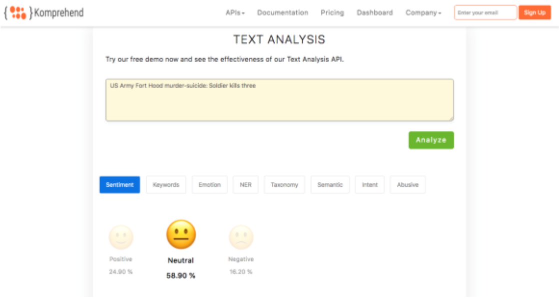
Subjectivity: **unknown**

Figure 3. uClassify

Figure 4. Text2Data

Meanwhile, Komprehend.io also shows a modern UI, an improvement over that of SentiStrength but using smiley faces to emphasize sentiment polarity. It should be noted, however, that the evaluation of the UI is out of the scope of this study. Figure 5, shown below, is the UI for Komprehend.io used to classify the 18th news headline from the data set, "US Army Fort Hood murder-suicide: Soldier kills three."

Note that in cases where the classification is for neutral polarity, the relative weights given to the positive and negative polarities of the item are taken into consideration. For example, in the case of the sample Text2Data classification in Figure 4 above, while the classification indicated is neutral, the score of +0.14 is used as basis to classify the same as positive polarity. A similar example is the one shown in Figure 5 at right. In this case, the classification is also neutral as indicated, but the item is more positive at 24.90% than it is negative at 16.20% so the classification is considered as positive for purposes of the metrics in this study.



TEXT ANALYSIS

Try our free demo now and see the effectiveness of our Text Analysis API.

US Army Fort Hood murder-suicide: Soldier kills three

Figure 5. Komprehend.io

RESULTS AND DISCUSSION

Sentiment analysis was conducted using the online SA-SaaS platforms Daniel Soper, SentiStrength, uClassify, Text2Data and Komprehend.io based on the 26 samples from [13], [14] and [15] used as test data set for this study. The sentiment analysis platforms were used to provide sentiment polarity classification for each of the test items. The confusion matrix for the results of the sentiment analysis using Daniel Soper is shown in Table 3. The confusion matrix for the results of the sentiment analysis using SentiStrength is shown in Table 4. The confusion matrices for the results of the sentiment analysis using uClassify, Text2Data and Komprehend.io are shown in Table 5, Table 6 and Table 7, respectively. Confusion matrices are the *de facto* measure of performance for classifiers. The diagonal in a confusion matrix show the correct classifications made. A cursory evaluation of the diagonal of the confusion matrices for the 5 SA-SaaS platforms show the classifications correctly made. Results showed SentiStrength and Komprehend.io leading with 20 correctly classified items, followed by uClassify and Text2Data, each with 16 and 15 correctly classified items. Daniel Soper is last with 8 correct classifications.

Table 3. Confusion Matrix for the Daniel Soper classification

		Manual Annotation		
		Positive	Negative	Total
Daniel Soper	Positive	4	9	13
	Negative	9	4	13
	Total	13	13	26

Table 4. Confusion Matrix for the SentiStrength classification

		Manual Annotation		
		Positive	Negative	Total
SentiStrength	Positive	10	3	13
	Negative	3	10	13
	Total	13	13	26

Table 5. Confusion Matrix for the uClassify classification

		Manual Annotation		
		Positive	Negative	Total
uClassify	Positive	7	6	13
	Negative	4	9	13
	Total	11	15	26

Table 6. Confusion Matrix for the Text2Data classification

		Manual Annotation		
		Positive	Negative	Total
Text2Data	Positive	7	6	13
	Negative	5	8	13
	Total	12	14	26

Table 7. Confusion Matrix for the Komprehend.io classification

		Manual Annotation		
		Positive	Negative	Total
Komprehend.io	Positive	9	4	13
	Negative	2	11	13
	Total	11	15	26

To evaluate the performance of the sentiment analysis done by each of the 5 platforms, the test metrics of accuracy, precision, recall or sensitivity, and specificity are calculated. As previously mentioned, accuracy is the measure of the ability of a system to correctly classify sentiments. The accuracy based on the results of the sentiment classification is shown in Table 8 for each of the sentiment classification as-a-service platforms. SentiStrength and Komprehend.io lead at both 76.92% accuracy. uClassify follows at 61.54% while Text2Data is almost a close fourth at 57.69%. The Daniel Soper platform placed last at 30.77%.

The second measure of performance is precision, which is indicative of the proportion of positive classifications that are actually correct. The precision based on the results of the sentiment classification is shown in Table 9 for each of the sentiment classification as-a-service platforms. SentiStrength inches forward at 76.92% ahead of its close competitor, Komprehend.io which recorded precision at 69.23%. uClassify and Text2Data are now tied at 53.85% while Daniel Soper lags behind at 30.77%.

Table 8. Accuracy of Sentiment Classification

Classifier	Accuracy
Daniel Soper	30.77%
SentiStrength	76.92%
uClassify	61.54%
Text2Data	57.69%
Komprehend.io	76.92%

Table 9. Precision of Sentiment Classification

Classifier	Precision
Daniel Soper	30.77%
SentiStrength	76.92%
uClassify	53.85%
Text2Data	53.85%
Komprehend.io	69.23%

The third measure of performance is recall or sensitivity. It is a measurement of the proportion of items that were labelled as positive that were correctly classified as such. The recall based on the results of the sentiment classification is shown in Table 10 for each of the sentiment classification as-a-service platforms. In terms of recall, Komprehend.io performed better and took the top spot at 81.82%, followed by close competitor SentiStrength at 76.92%. uClassify also inched ahead of its competitor at 63.64% while Text2Data only did 58.33%. Daniel Soper remained the least performing at 30.77%.

The final measure of performance is specificity, which is similar to recall or sensitivity, only that it looks at the negative polarity instead of the positive. It is a metric of how much of the items expected to be classified as negative were actually classified as negative. The specificity based on the results of the sentiment classification is shown in Table 11 for each of the sentiment classification as-a-service platforms. SentiStrength leads with 76.92%, followed by Komprehend.io at 73.33%. uClassify ranks third with 60.00% and Text2Data comes next with 57.14%. Finally, Daniel Soper is at 30.77%.

Table 10. Recall of Sentiment Classification

Classifier	Recall
Daniel Soper	30.77%
SentiStrength	76.92%
uClassify	63.64%
Text2Data	58.33%
Komprehend.io	81.82%

Table 11. Specificity of Sentiment Classification

Classifier	Specificity
Daniel Soper	30.77%
SentiStrength	76.92%
uClassify	60.00%
Text2Data	57.14%
Komprehend.io	73.33%

The four metrics used to quantify performance of the different SA-SaaS platforms in this study measure different aspects of how good the classifier performed the classification. Not one of them is better than the others, considering that each metric evaluates a different aspect of performance. However, if each metric is to be considered of equal importance in a weighted criteria set, SentiStrength would emerge on top with an average performance metric of 76.92%. Komprehend.io follows close at 75.33%. uClassify would come third at 59.76% average, followed by Text2Data at 56.75%. Daniel Soper comes last at 30.77%. It should be noted, however, that Daniel Soper is quite stable with its classification performance across all metrics. Similarly, SentiStrength is likewise very stable, with its classification performance varying only very slightly across the metrics. This is shown in Figure 6 below, a bar graph that shows the performance of each SA-SaaS platform across the defined metrics.

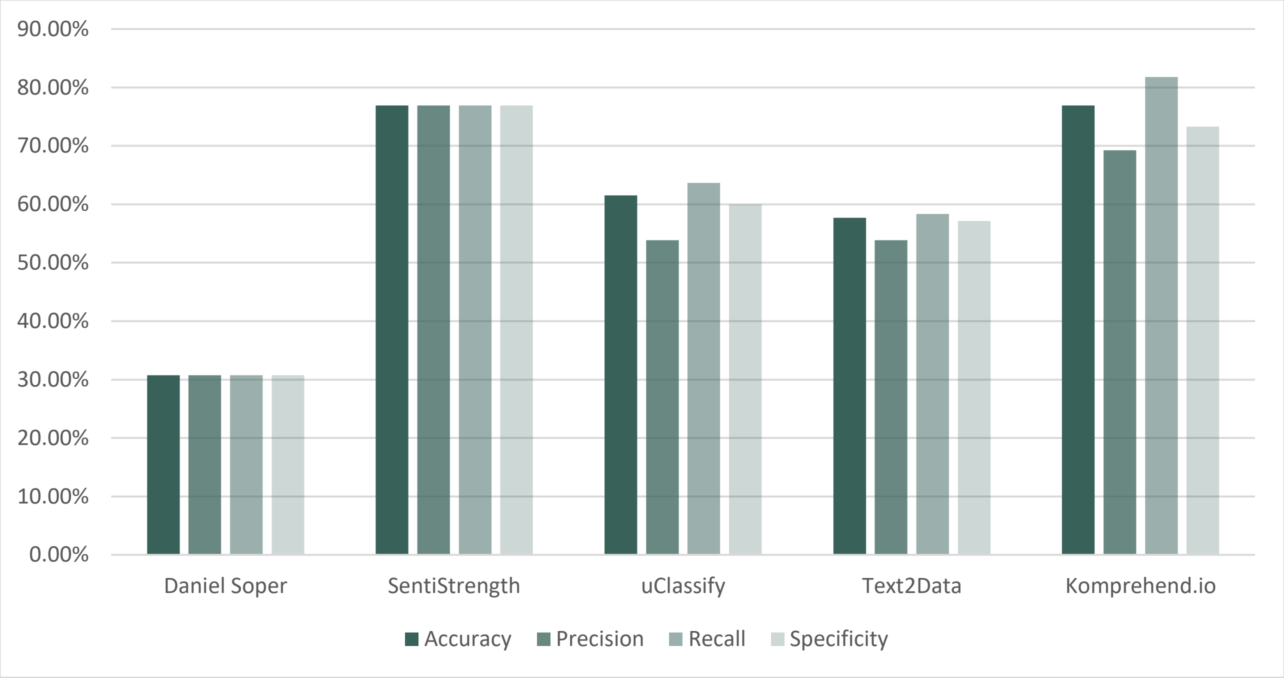


Figure 6. Bar chart showing performance of the classifiers across all four metrics.

CONCLUSION AND RECOMMENDATIONS

A standard data set annotated with polarity information is a good input to test the performance of black box SA-SaaS. In this study, two platforms, SentiStrength and Komprehend.io, performed really well across all metrics and at levels that, considering inter-rater differences, are very close to how humans would have made the same classification. When selecting a SA-SaaS platform, the best candidate is the one that can identify sentiment polarity correctly more than the others. Of those that can identify sentiment polarity correctly more than the others, the best choice would be those that can identify sentiment polarity consistently more than the others. Results from this study show that SentiStrength identifies sentiment polarity correctly more than the others, it also does the job quite consistently.

Faced with black box SaaS, it would be useful to perform a similar analysis, ideally with a bigger data set for a more statistically significant sampling. The size of the data set, of course, will be limited by the work that needs to be done when using a SA-SaaS platform with a web UI and the number of trial runs allowed for demonstration versions if the model used in this study is to be followed. That said, developing a data set for the purpose of making it readily available to other researchers in sufficient volume would be a plus.

REFERENCES

[1] S. Mohammad, B. Dorr, and C. Dunne, “Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*: Singapore; pp. 599-608, 2009, https://aclanthology.org/D09-1063.pdf.

[2] B. Liu, “Sentiment Analysis and Subjectivity,” *Handbook of Natural Language Processing* 5, 1–38, 2010, https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf.

[3] A. Esuli and F. Sebastiani, “SentiWordNet: A publicly available lexical resource for opinion mining,” *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*: Genoa, Italy; pp. 417-22, 2006.

[4] Y. Mejova, “Sentiment Analysis: An Overview,” *unpublished comprehensive exam paper at the Computer Science Department, University of Iowa*, 2009, https://www.academia.edu/291678/Sentiment_Analysis_An_Overview.

[5] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. Subrahmanian, “Sentiment analysis: Adjectives and adverbs are better than adjectives alone,” *Proceedings of International Conference on Weblogs and Social Media, ICWSM*. Boulder, CO, 2007.

[6] Y. Lu, X. Kong, X. Quan, W. Liu and Y. Xu, “Exploring the Sentiment Strength of User Reviews,” L. Chen, C. Tang, J. Yang, and Y. Gao (eds), *Web-Age Information Management, WAIM 2010, Lecture Notes in Computer Science*, Vol. 6184, 2010, Springer, Berlin, Heidelberg, doi:10.1007/978-3-642-14246-8_46.

[7] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, “Analysis of Political Sentiment Orientations on Twitter,” *Procedia Computer Science*, 167: 1821-1828, 2020, doi: 10.1016/j.procs.2020.03.201.

[8] M. Bansal, C. Cardie, and L. Lee, “The power of negative thinking: Exploiting label disagreement in the min-cut classification framework,” *Proceedings of the International Conference in Computational Linguistics (COLING)*: Manchester UK; pp. 15-18, 2008, https://aclanthology.org/C08-2004.pdf.

[9] N. Diakopoulos, “Algorithmic accountability reporting: On the investigation of black boxes,” *A Tow/Knight Brief*. Tow Center for Digital Journalism, Columbia Journalism School, Columbia University, 2014, doi: 10.7916/D8ZK5TW2

[10] J. Mihaljević, J., “Analysis and Creation of Free SentimentAnalysis Programs,” *Croatian Journal for Journalism and the Media*, 25(1): 83-104, 2019, doi: 10.22572/mi.25.1.4.

[11] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, “Sentiment analysis: A review and comparative analysis of web services,” *Information Sciences*, 311: pp. 18-38, 2015, doi: 10.1016/j.ins.2015.03.040.

[12] M. Thelwall, K. Buckley, G. Paltoglou, and D. Cai, “Sentiment strength detection in short informal text,” *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558, 2010, http://www.scit.wlv.ac.uk/~cm1993/papers/SentiStrengthPreprint.doc. [14] J. Reis, F. Benevenuto, P. O. S. Vaz de Melo, R. Prates, H. Kwak, and J. An, “Breaking the News: First Impressions Matter on Online News,” *The International AAAI Conference on Web and Social Media (ICWSM)*, 2015, https://www.researchgate.net/publication/284162897_Breaking_the_News_First_Impressions_Matter_on_Online_News.

[13] S. Lawrie, “Newspaper coverage of psychiatric and physical illness,” *The Psychiatrist*, 24:104-106, 2000, https://www.researchgate.net/publication/247805527_Newspaper_coverage_of_psychiatric_and_physical_illness.

[14] J. Reis, F. Benevenuto, P. O. S. Vaz de Melo, R. Prates, H. Kwak, and J. An, “Breaking the News: First Impressions Matter on Online News,” *The International AAAI Conference on Web and Social Media (ICWSM)*, 2015, https://www.researchgate.net/publication/284162897_Breaking_the_News_First_Impressions_Matter_on_Online_News.

[15] H. Nisar and E. Bleich, “Group status, geographic location, and the tone of media coverage: Jews and Muslims in New York Times and Guardian Headlines, 1985–2014,” *Comparative Migration Study*, 8(3), 2020, doi:10.1186/s40878-019-0153-3.