

May 16, 2021

Quinn E Knudsen

IST 736

Syracuse University

(E)Motion Picture

An Evaluation of MTurk in Providing Sentiment Classifications on Movie Reviews

**Table of Contents**

**Introduction**……………………………………………………………………………..pg 2

Areas of Focus ………………………………………………………………......pg 2

Movie Reviews……. ...……………………………………………………….....pg 2

Sentiment Analysis …………………………………………………………….. pg 3

**Analysis**………………………………………………………………………………..…pg 4

About the Data ...…………………………………………………………….…...pg 4

Data Cleaning & Prep..……………………………………………………….…..pg 4

Sentiment Classifiers …………..…………………………………………….…..pg 5

**Results**…....………………………………………………………………………….…..pg 5

Analysis 1: Sentiment Classification of Text..…………………………….……..pg 5

*Research Question 1………*………………………………………………pg 5 *Research Question 2………*………………………………………….….. pg 8 *Research Question 3………*………………………………………….….. pg 9

**Conclusions**………………………………………………………………………….…..pg 10

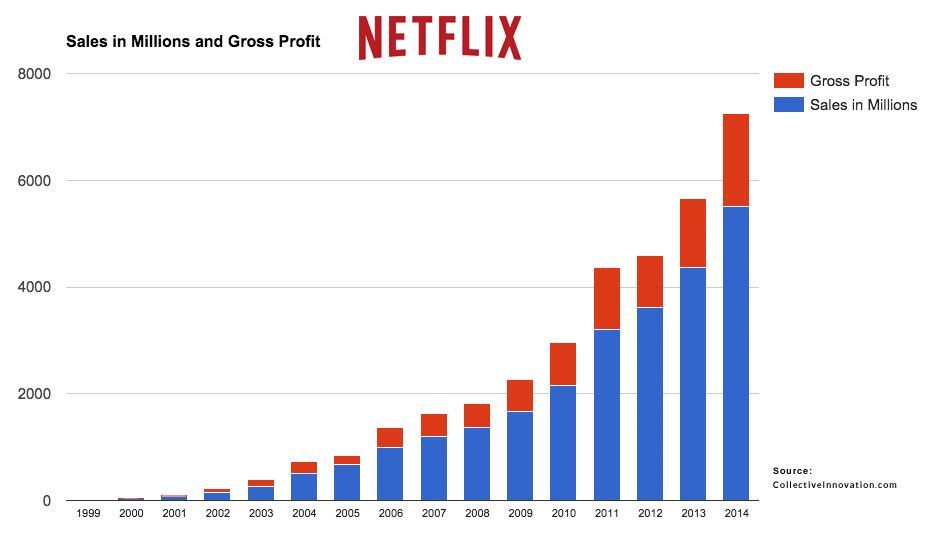
**Introduction**

**Areas of Focus**

The scope of this paper is to explore the utility of using Amazon MTurk as a resource for machine learning dataset construction. In this instance, the specific interest is around classifying text as positive or negative. This task will ask respondents to provide a classification for a movie review dataset containing three positive and three negative reviews. These reviews will be scored against machine learning sentiment classifiers, peer classifiers and author categorization.

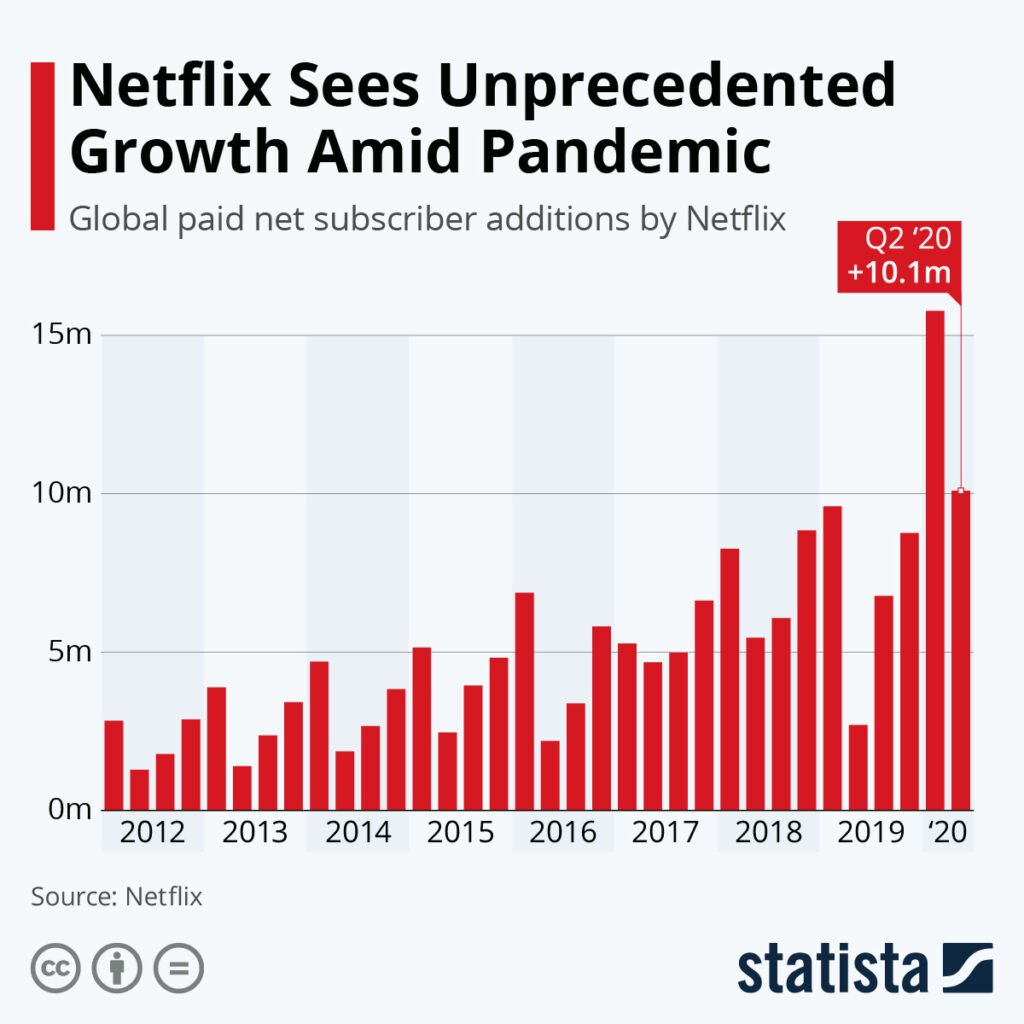
**Movie Reviews**

With the emergence of Netflix, Hulu, Amazon Prime and other stream services, it might be the wrong time to take a long-term investment position in Blockbuster. However, it is a great time to be a movie enthusiast with an all-time level of convenience and accessibility to one’s favorite shows, actors and directors. In fact, the growth of global video streaming such as amazing have grown exponentially as seen in Figure 1 below.



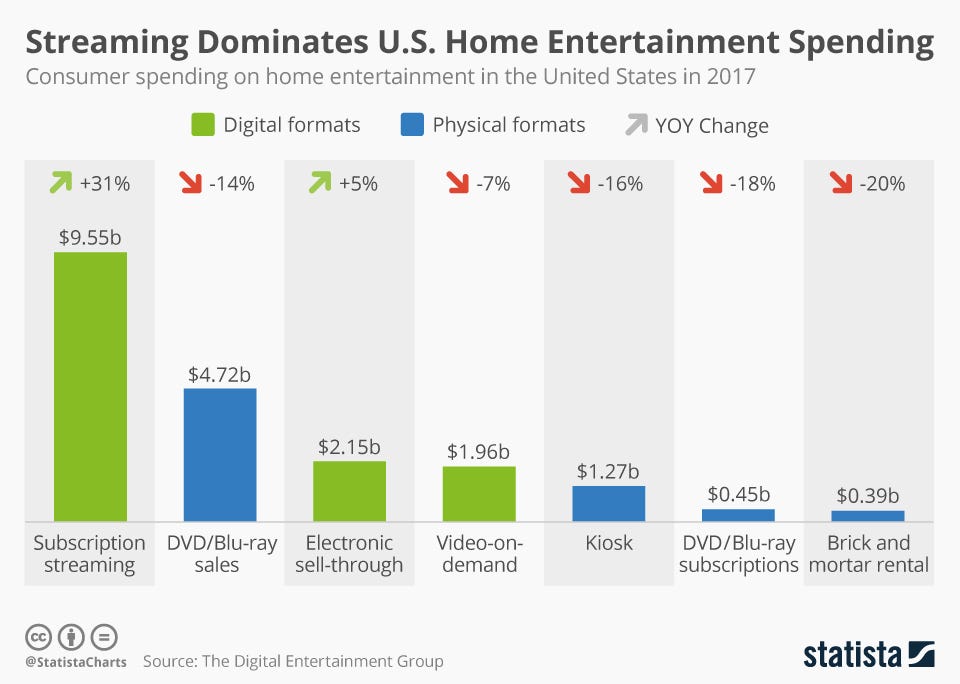
*Figure 1.* Growth of Netflix sales and gross profit from 1999-2014.

In recent years, amplified by the global pandemic, the demand for movie streaming services has only been amplified further. A growing number of traditional movie or television companies such as Disney and CBS now offer streaming subscription services online. Despite the growing market competition, Netflix remains the cream of the crop with market share and has continue to grow as noted in Figure 2.



*Figure 2.* Netflix dominance in streaming services has persisted in recent years.

Just as the movies have shifted over the years to digital providers, so have the manner in which movie goers seeks out ratings. Movie reviews are posted online and can even feed algorithms designed to suggest movies to online consumers. Scraping reviews for meaning can be an arduous but reward task. Companies that leverage traditional numeric ratings and also text analytics to understand the sentiment behind the movie watcher can create a competitive advantage in the marketplace.



*Figure 3.* Home entertainment spending by medium in the United States in 20117.

**Sentiment Analysis**

One applicant of machine learning popularized in recent years is text analytics-or the process of converting large amounts of unstructured written human linguistics into quantitative data capable of uncovering insights meaningful for human interpretation. Sentiment analysis is a specific domain of natural language processing that tries to determine whether the written information is positive, negative or neutral based upon the polarity of the words used in the text. This can prove to be a tall order due to the complexities of language including harder to detect nuances such as sarcasm, slang, negation or amplification. The partial focus of the analysis portion of this paper will explore an applicant of sentiment analysis on commentary related to movie reviews.

**Analysis**

**About the Data**

The dataset contains a series of written statements about movies to test the interrater reliability amongst Amazon MTurk reviewers and friends. The focus of this paper is on the efficacy of using Amazon MTurk as a data collection tool specific to the task of classifying sentences.

The initial dataset contained two columns: a statement about a movie and a sentiment rating. Six statements were written by the author with three being classified as positive and as negative. Four of the statements were intentionally written so that they would be easy to classify, while two statements were written in a more ambiguous or sarcastic manner that might present a challenge to raters.

These statements were then analyzed with three Python sentiment classifiers (VADER, TextBlob and AFINN). The algorithm classifications, MTurk classifications, and friend classifications were then stored in a data frame and tested for interrater agreement.

**Analysis: Pseudo Movie Reviews**

**Data Cleaning & Prep**

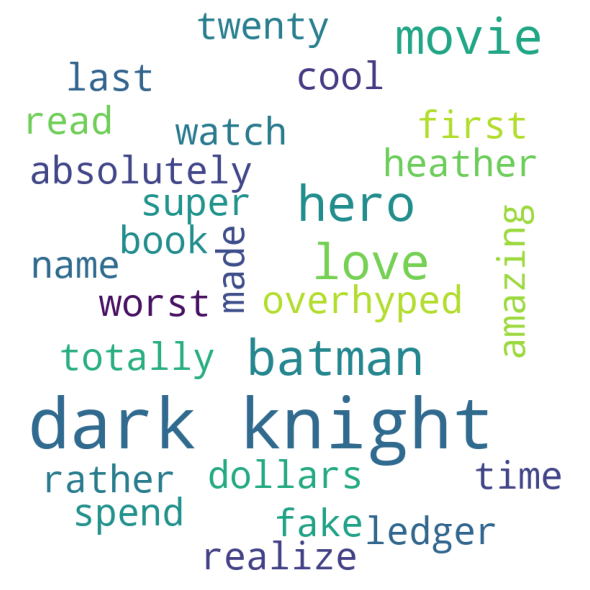
The author generated a list of six movie reviews with perfect classification balance between positivity and negativity. As such, the quality of the data was disproportionality clean and required no data cleansing in preparation for sentiment classification. Four of the statements were written in a manner that would be easy to classify using overt sentiment clues, while two sentences were written with a bit more ambiguity. A full list of the raw data can be found below in Table 1.

|  |  |
| --- | --- |
| **Sentiment** | **Review** |
| **Positive** | I love the Dark Knight! Heather Ledger was absolutely amazing. |
| **Positive** | Batman is such a cool hero. I love the movie and how he is super. |
| **Positive** | I would spend the last twenty dollars to my name to watch the Dark Knight again for the first time. |
| **Negative** | The Dark Knight is the worst movie ever. Totally overhyped. |
| **Negative** | Batman is a fake hero. I do not like him at all or the Dark Knight. |
| **Negative** | The Dark Knight made me realize I would rather read a book. |

*Table 1.* Statements made about the Dark Knight in their raw form with expected classification.

*Word cloud*

Data preparation was needed to generate a word cloud. This was achieved through tokenization of the text documents and the removal of common English stop words (from the nltk library) and custom stop words “go” and “going”. Figure 4 demonstrates the most common words used by the author with a clear focus on Dark Knight, Batman, love and hero.



*Figure 4.* Depiction of the most common words generated by the authors.

**Sentiment Classifiers**

*TextBlob*

TextBlob uses API for common NLP tasks such as sentiment analysis by assigning polarity (positive, negative or neutral) to a sentence. This classifier was used to compare with the ensuing sentiment classifiers as well as to compare sentiment written by the two authors. Sentiment with a polarity > 0 was considered to be positive, equal to 0 neutral and < 0 negative.

*Afinn*

The AFINN lexicon is a list of English terms manually rated for valence with an integer between -5 (very negative) and +5 (very positive) Årup Nielsen between 2009 and 2011 (<http://corpustext.com/reference/sentiment_afinn.html>).

*Vader*

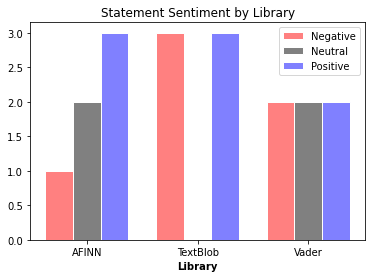
Vader is a lexicon rule-based sentiment classification tool adept at handling social media and other forms of text. Vader uses a sentiment analyzer to score each text document with sensitivity to polarity (positive or negative) and intensity. While Vader provides scores for negativity, neutrality and positivity, the comparison methodology chosen for this analysis was to compare the compound score. The compound scores each word in the lexicon and then normalized the result between -1 (highly negative) and +1 (highly positive).

**Results**

**Analysis 1: Sentiment Classification of Text**

*Research Question 1:* Can Python sentiment classifier correctly label movie reviews?

Each sentiment classification was recoded so that positivity was indicated by a value > 0, neutrality = 0 and negativity < 0. For the AFINN, TextBlob and Vader classifications overall positivity, neutrality and negativity were determined. As exemplified in the comparison of Figure 9, negative classification was the least likely categorization. There was variation in that Vader and AFINN classified with neutrality. TextBlob did not classify the text as neutral and equally split positive and negative. AFINN under classified negativity.



*Figure 5.*  Polarity of statements for three Python sentiment classifiers for all statements written.

Utilizing the Python sentiment classifiers, each statement was scored so that and compared with the predetermined sentiment. As expected, the more ambiguous sentences scored the lowest with 33% accuracy for each. TextBlob correctly categorized each sentence. The largest blunder was AFINN’s misappropriation of sentence 5 as positive when it is in fact negative.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Text** | **TextBlob** | **AFINN** | **Vader** | **Accuracy** |
| 1 | I love the Dark Knight! Heather Ledger was absolutely amazing. | positive | positive | positive | 100% |
| 2 | Batman is such a cool hero. I love the movie and how he is super. | positive | positive | positive | 100% |
| 3 | I would spend the last twenty dollars to my name to watch the Dark Knight again for the first time. | positive | neutral | neutral | 33% |
| 4 | The Dark Knight is the worst movie ever. Totally overhyped. | negative | negative | negative | 100% |
| 5 | Batman is a fake hero. I do not like him at all or the Dark Knight. | negative | positive | negative | 66% |
| 6 | The Dark Knight made me realize I would rather read a book. | negative | neutral | neutral | 33% |

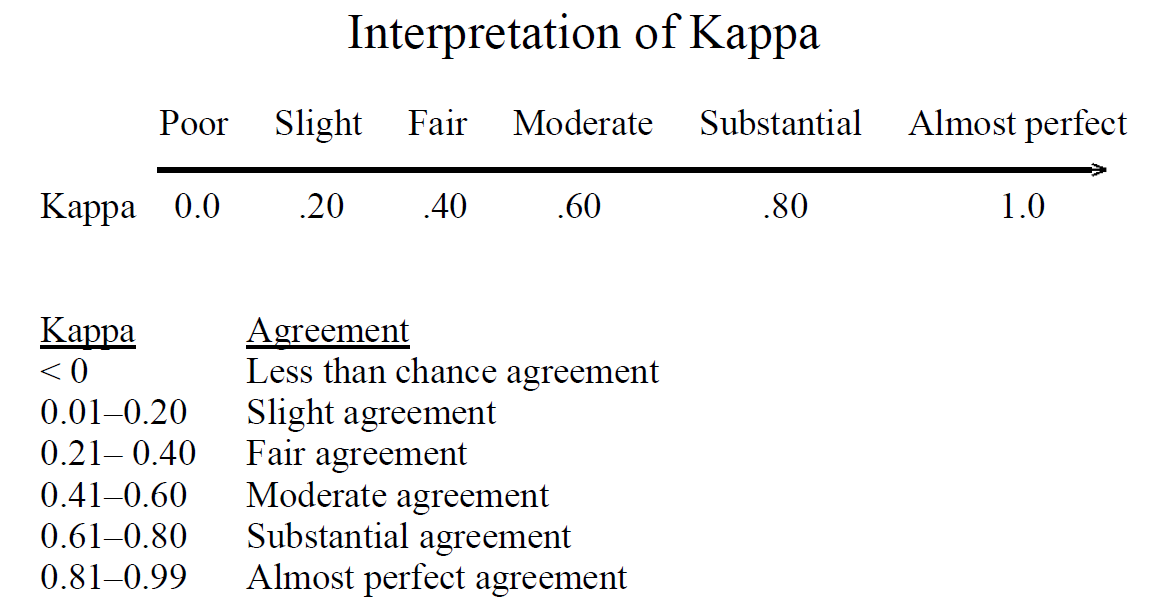
*Table 2.*  Sentiment classification for the Python packages.

Agreement between classifiers was also tested as depicted in Table 1 below with the highest agreement between AFINN & Vader at 85%. This may partially be explained by the lower negativity found with TextBlob. The 15 most agreed upon positive statements and the 15 most agreed upon negative statements were then converted into a corpus for analysis 2.

|  |  |
| --- | --- |
| **Classifiers** | **Cohen’s Kappa** |
| **AFINN & TextBlob** | .5 |
| **AFINN & Vader** | **.75** |
| **TextBlob & Vader** | .25 |
| **All** | .5 |

*Table 3.*  Rate of agreement between classifiers finds AFINN and Vader to be the most similar.

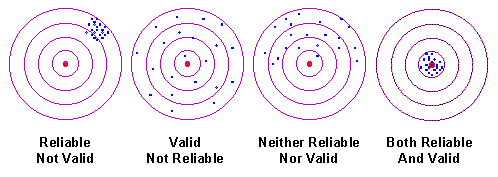
The Cohen’s Kappa score achieved here would fall into the moderate agreement range leaving clear room for opportunity to improve.



*Figure 6.*  Interpretation of Cohen’s Kappa.

A calculation of Krippendorff’s alpha found similar results with an alpha of .59. Values range from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement. Krippendorff suggests: “[I]t is customary to require α ≥ .800. Where tentative conclusions are still acceptable, α ≥ .667 is the lowest conceivable limit (Krippendorff 2004, p. 241).” Therefore, the machine classifiers would not meet the desired thresholds for interrater agreement.

Beyond agreement and interrater reliability, another key factor to consider is the rate of accuracy with the classification task. While it is great to have reliability, as demonstrated visually in the far left target in Figure 7 below, reliability without accuracy is devoid of value. It is crucial to consider the validity or in this case the accuracy of the classifications with the optimal performance exemplified in the far right target.



*Figure 7.* The dart board example of reliability and validity. Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. <https://www.socialresearchmethods.net/kb/>.

*Research Question 2:* Can peers of the author correctly label the sentiment of movie reviews?

In order to test the second research question, six associates were tasked with classifying the same six sentences as positive negative or neutral. It was found that for five of the six statements there was complete agreement. For the statement, “The Dark Knight made me realize I would rather read a book.”, there was much debate. Each rating was independent and found 50% in favor of negative and 50% in favor of neutral. The results for these evaluations can be found in Table 4 below.

|  |  |  |
| --- | --- | --- |
| **ID** | **Text** | **Correct Assignment** |
| 1 | I love the Dark Knight! Heather Ledger was absolutely amazing. | 100% |
| 2 | Batman is such a cool hero. I love the movie and how he is super. | 100% |
| 3 | I would spend the last twenty dollars to my name to watch the Dark Knight again for the first time. | 100% |
| 4 | The Dark Knight is the worst movie ever. Totally overhyped. | 100% |
| 5 | Batman is a fake hero. I do not like him at all or the Dark Knight. | 100% |
| 6 | The Dark Knight made me realize I would rather read a book. | **50%** |

*Table 4.*  Sentiment assignment for six peer reviewers demonstrates inconsistency on one item.

Items 3 and 6 were intended to be more challenging to classify. Item 6 shows that classification is a subjective art and relies upon individual perception. However, is algorithms evolve over time, this also shows the potential for algorithms to outpace human ability to classify sentiment.

*Research Question 3:* Will MTurk reviews differ from peer reviews?

In an effort to evaluate the effectiveness of MTurk as a classification resource, the same statements were uploaded to Amazon MTurk. Participants were offered $.10 in exchange for each statement classified. 5 turkers were recruited and within five minutes of posting the job it was completed. The average completion time was 25 seconds per item with a range of 5 to 51 seconds. Language requirements were English and past performance was not required in order to assure tasks completion. Utilizing verified workers would be advisable with a larger budget to control quality and spam. English speaking workers are needed because the items were written in English. Any respondent that demonstrated abnormal response times or clear wild guessing will be removed to preserve data quality and integrity. Efficiency of collect was key, so it was decided to not over stipulate credentials for labeling and place a general faith in humanity for honesty and effort.

Table 5 demonstrates the results by item were poorer than peer reviews, but slightly better than Python sentiment analyzers. It was clear that negative statements were more challenging for the turkers to accurately label.

|  |  |  |
| --- | --- | --- |
| **ID** | **Text** | **Correct Assignment** |
| 1 | I love the Dark Knight! Heather Ledger was absolutely amazing. | 100% |
| 2 | Batman is such a cool hero. I love the movie and how he is super. | 100% |
| 3 | I would spend the last twenty dollars to my name to watch the Dark Knight again for the first time. | 80% |
| 4 | The Dark Knight is the worst movie ever. Totally overhyped. | 60% |
| 5 | Batman is a fake hero. I do not like him at all or the Dark Knight. | 60% |
| 6 | The Dark Knight made me realize I would rather read a book. | 60% |

*Table 5.*  Sentiment assignment for six peer reviewers demonstrates inconsistency on one item.

In total, the budget of $3.60 was used. No one turker could be identified as a spammer based upon response time average and accuracy, however it is expected with larger classification tasks spammers would infiltrate. With this small assignment, it was hard to clearly identify a spammer.

*Research Question 4: Which classification group will have the highest KPIs?*

Below Table 6 demonstrates the overall classification accuracy as well as the average number of mislabeled statements. The best rating group would be the peer raters with the highest marks on accuracy, Cohen’s kappa and Krippendorff’s alpha. Interestingly, while the Python classifiers were slightly less accurate, they had higher interrater reliability scores than the five turkers. This indicates further evidence that controlling for spammers is crucial in large scale labeling tasks in order to weed out wild low attention labelers and bots. The turkers were less consistent in their categorization, but provide a cheap alternative to a team trying to tackle millions of reviews.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Mislabeled Statements** | **Accuracy** | **Average Cohen's Kappa** | **Krippendorff's Alpha** |
| Python Packages | 1.7 | 72% | 0.5 | 0.59 |
| Peers | 0.5 | 92% | 0.83 | 0.95 |
| MTurk | 1.4 | 77% | 0.24 | 0.33 |
| Average Model | 1.2 | 80% | 0.52 | 0.62 |

*Table 6.*  KPIs for various rater metrics.

Taking the 14 classified sentences, aggregate accuracy scores were calculated as displayed in Table 7. Overall, positive statements were more accurately labeled than negative statements and the lowest scoring sentences for each polarity were the challenge sentences (3 and 6). This confirmed the hypothesis that the challenge sentences would have the lowest accuracy, however difficulty by sentiment was not anticipated.

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Text** | **Correct Assignment** | **Polarity Accuracy** |
| 1 | I love the Dark Knight! Heather Ledger was absolutely amazing. | 100% | **93%** |
| 2 | Batman is such a cool hero. I love the movie and how he is super. | 100% |
| 3 | I would spend the last twenty dollars to my name to watch the Dark Knight again for the first time. | 79% |
| 4 | The Dark Knight is the worst movie ever. Totally overhyped. | 86% | **71%** |
| 5 | Batman is a fake hero. I do not like him at all or the Dark Knight. | 78% |
| 6 | The Dark Knight made me realize I would rather read a book. | 50% |

*Table 7.*  Results summary for all sentences labeled by the 14 classifiers.

**Conclusion**

As the demand for online web services for movies grows, so too does the utilization online movie reviews. As the abundance of data grows, the value of extrapolating meaning from the textual reviews does as well. Understanding public sentiment on a movie or show can provide a competitive advantage for streaming services hoping to strategically stream the best content.

Just humans tend to disagree about their feelings over a popular movie, Python sentiment classifiers were not in agreement about the reviewer’s sentiment. Findings suggest that within the practice of analyzing text for sentiment, there is variable agreement amongst the text classifiers. Depending upon the Python package used the sentence may be classified as positive, neutral or negative.

It was also demonstrated that peer classifiers with a nonexpert background tend to disagree on statements that are less overt. Unlike Python, none of the classifications were found to be in the oppositive direction, however peer reviewers used the neutral option on one particularly tricky item. With subjectivity involved, it presents a challenge in making uniform decisions on sentiment classification.

As it relates to the efficacy of using Amazon MTurk, somewhat lukewarm results were found. While it is certainly a viable solution for gathering training labels at scale, precautions should be taken to avoid hasty responding, bots or misaligned labelers. Things to keep in mind are geographic location and language and it is good to monitor response times and accuracy to remove poor data quality provided by the labelers. It would be wise, albeit with a cost, to pay the extra money to get verified (vetted) labelers in MTurk to provide the more optimal labeling. Given a budget, this could be a viable solution for labeling, however with extremely large datasets or with data that will need frequent labeling, the cost at scale should be considered before allocations sneak up.

Ultimately, companies should consider using turkers to classify movie reviews as the upfront cost may be offset by a competitive advantage in understanding the preferences gleaned from text-based movie reviews. The resources needed to understand millions of movie reviews without automated labeling would be an impossible task without scalable solution.