Sentiment Analysis of Stomp

Jiang Zhiheng (4S109), Tan Yi Kai (4S123), Huang Junwei (4S213)

Group 09-04

Hwa Chong Institution

Sentiment Analysis of Stomp

Jiang Zhiheng[1], Tan Yi Kai[1], Huang Junwei[1]

[1]Hwa Chong Institution, 661 Bukit Timah Road, Singapore 269734

Abstract

Citizen journalism is the new vogue. As media becomes democratised, many citizens have begun to use online platforms to voice their opinions on current affairs. The rise of the Internet has given everyone a microphone, amplifying their opinions through the Web. Stomp.sg is Singapore's most widely used social networking and citizen media website, and the opinions voiced there can be an insightful cross-section into the views of ordinary Singaporeans. We aim to analyse the sentiments of these opinions on a larger scale, as the data obtained could help politicians and public servants to gain valuable insight into the lives and views of Singaporeans, better serving the people's interests.

Introduction

The rise of the Internet gave birth to online forums, where many netizens participated in various discussions that can cover a broad range of topics, including everything from academics to politics (Holtz, Kronberger and Wagner, 2012). Locally, the most popular online forum is the Straits Times Online Mobile Print, abbreviated as Stomp, with more than 66.8 million monthly page views in 2014. These forums provide a platform to aspiring citizen journalists. These are ordinary citizens, with no professional journalistic training, writing about news stories that concern them and their communities. Essentially, it is journalism that every person can engage in, with little prerequisites to entry. (Noor, 2017). Citizen journalism is also an invaluable medium to disseminate information that professional journalists cannot access, such as in war-torn areas. Citizen journalism can contribute to increased civic engagement (Pendry and Salvatore, 2013) and provide crucial firsthand insights that are unavailable through traditional news media (Martinsson, 2009), such as in Kashmir in 2008 where local residents detailed the violence and atrocious war crimes committed by the paramilitary forces and helped to garner international support (Noor, 2017). This was information beyond the reach of traditional journalists. Such insights can also improve governance, as they give the government vital information that might not have been available through traditional news sources, helping them make decisions (Martinsson, 2009). To help analyse the citizen journalism articles, we have decided to use sentiment analysis algorithms, due to their proven efficiency and accuracy on similar platforms such as

Twitter (Sarlan, Nadam and Basri, 2014). We have also adopted the Naive Bayes algorithm to intelligently classify the stories into four main categories, as the accuracy of the Naive Bayes algorithm increases with larger input sizes (Sanaz, 2012). The results of our analysis will be presented via Tableau for possible use by government agencies and academic scholars. From observing the trend in sentiments against number of views and comments, we will also be able to predict the general reaction of Stomp users to articles of a certain sentiment.

Literature Review

Noor, R. (2016). Citizen Journalism vs. Mainstream Journalism: A Study on Challenges Posed by Amateurs. *Athens Journal of Mass Media and Communications,3*(1), 55-76. doi:10.30958/ajmmc.3.1.4

This paper performs a comparative study on professional and citizen journalism, concluding that citizen journalism poses little threat to professional journalism as it is still in its formative years. Instead, the paper proposes that the two might be complementary, with each having its unique advantages. The paper focuses on a study done in Kashmir, a region on the India-Pakistan border prone to regular military conflict. One interesting insight provided is how citizen journalists can write news stories that professional journalists have no access to. The paper cites a paramilitary incident in Kashmir to justify its stand. Citizen journalists blogged extensively about the war crimes of the paramilitary groups and quickly garnered international support. However, the incident was not first reported in any newspaper or media outlet as professional journalists, considering that Kashmir is politically unstable, may hesitate to travel there to report, increasing the importance of the role of citizen journalism in society to report on issues that may be otherwise neglected by the mainstream media.

Sarlan, Aliza & Nadam, Chayanit & Basri, Shuib. (2014). Twitter sentiment analysis. 212-216. 10.1109/ICIMU.2014.7066632.

A piece of research conducted by Sarlan et al in 2014 reported on the methods and results of sentiment analysis on Twitter, a similar English-language online social media and commentary platform. The results shows that sentiment analysis is capable and efficient dealing with large amounts of data, with time complexities around O(N) and acceptable runtimes, but this could be attributed to Twitter limiting every post to just 140 characters. It also accepts common short forms and contractions. However, it cannot recognise misspelled words which are very common on such platforms or emoticons,

symbols which are commonly used to convey emotions. Additionally, it did not have to deal with Singlish, the Singaporean variant of English that involves many slang words.

Sanaz. (2012). Online Forums Hotspot Prediction Based on Sentiment Analysis. *Journal of Computer Science,8*(8), 1219-1224. doi:10.3844/jcssp.2012.1219.1224

Similar to the previous study, this study also explores methods of classification of hotspots on Internet forums. However, in addition to the K-means clustering method explored by the previous study, it also considers other approaches, a decision tree algorithm J48 and the Naive-Bayes algorithm. On a English-language forum forums.digitalpoint.com, all 3 algorithms have achieved comparable success, though there are slight differences in performance. K-means clustering performs best (~90% accuracy) when K=5, while the accuracy of J48 and Naive-Bayes increases with increasing K. This opens up different approaches that cater to different input sizes of the same problem type.

Data

Variable	**Method of Collection**
Total No. of Stories Analysed	899
Duration of Analysis	18 April 2019 to 12 July 2019
No. of Views	Extracted directly from Stomp.sg (6 <= n <= 88256)
No. of Comments	Extracted directly from Stomp.sg (0 <= n <= 258)
Sentiment of Posts	Computed the compound sentiment of all the words in the article using VADER Sentiment Analysis
Sentiment of Comments	Computed the average compound sentiment of individual comments by processing of all the words in each comment using VADER Sentiment Analysis (Hutto and Gilbert (2014)).

<table>
<tr><td>Sentiment of Title</td><td>Computed the compound sentiment of all the words in the title using VADER Sentiment Analysis</td></tr>
<tr><td>No. of Words</td><td>Counted the number of words in the article</td></tr>
<tr><td>Category</td><td>Collected the tags of each article before assigning the respective category. The categories are assigned as follows:

<table><tr><td>Category</td><td>Tags</td></tr><tr><td>Inspiring Stories</td><td>heartwarming, kudos, inspiring, kind, helpful, hero, heroes</td></tr><tr><td>Crime</td><td>police, arrest, courts and crime, crime, court, murder, theft, abuse, stealing, death, stab, stealing, arrested, illegal, courtroom, fraud, cheating, appeal, scam, extortion, death, jailed, body, raid, rubbish, littering, litter, charged, vandalism, rape, threaten</td></tr><tr><td>Accidents</td><td>accident, fatal, death, pedestrian, injured, tragic attack, dispute, road rage, fire, damage, hit and run, dangerous, safety, scdf, fall, tree, fallen tree, motorcyclist, car accident</td></tr></table></td></tr>
</table>

	Controversy	inconsiderate, expensive, weird, unusual, disgusting, delivery, controversy, controversial, dirty, messy, bad service, inappropriate, viral, youths behaving badly, unusual sight, rude
Location	Extracting articles mentioning each planning neighbourhood in Singapore based on search	

Methods

1. Data Collection
    a. We scraped 899 articles from Stomp from 18 April 2019 to 2 August 2019 using Python's BeautifulSoup module, collecting key data including tags, number of views, comments, number of comments, date of publication.
    b. To collect location data, we did a Stomp Search for every planning neighbourhood in Singapore and we parsed the first 10 articles for each location if available, and computed all of the above domains of research.
    c. Article parsing
        i. We first replaced Singlish words with their English equivalent.
        ii. We then corrected spelling errors through the SymSpell algorithm , that initialises a hashtable of strings that contains all variants of words that have one missing letter.

            After checking a word against a dictionary and confirming that it has an error, all possible variants of that word with one missing letter and the original word is looked up on the hashtable.

            This allows spelling errors to be corrected up to 1 deleted character and 1 extra character. When no possible variants can be found, we assume the word to be a proper noun, and leave it as it is.

    d. We processed the sentiments of the posts and comments using the 'compound' value using VADER Sentiment Analysis and Python 3.6. The range of possible values is from -1 to 1, with the negative end representing negative sentiments and the positive end representing positive sentiments.
    e. We saved this data into a Microsoft Excel sheet using a Python Script.
2. Data Visualisation:
    a. Using Tableau, we visualised the data by considering many domains, including the following:
        i. No. of Views
        ii. No. of Comments
        iii. Sentiment of Posts
        iv. Sentiment of Comments
        v. Sentiment of Title
        vi. No. of Words
        vii. Category (Identified based on tags as mentioned earlier)
        viii. Date and Time
        ix. Location
    b. Location, No. of Words and Sentiment of Title were not included in our results analysis below as no significant trends were deduced.
3. Classification
    a. We trained a Naive-Bayes Classifier on a list of articles and their respective categories using NLTK by assigning a probability to each category for each article based on the word choice of the article, as shown in Fig 1.1.
        i. The categories are Inspiring Stories, Crime, Accidents and Controversy. These categories were determined after examining the tags found on Stomp. As the tags can be rather inconsistent, we chose to limit to 4 categories for accuracy.
    b. We trained on 440 articles and tested on 110 articles, after randomising the order of articles to increase accuracy.
    c. We then tested our classifier on a new list of articles and we found that we had an accuracy of over 70%.
4. Prediction
    a. We determined the category of a given article using our Naive Bayes Classifier mentioned above.
    b. We calculated the average comment sentiment and number of comments of articles in each category.

c. We used the average comment sentiment and number of comments to determine whether the article will be well received by the public, and how popular/unpopular it will be.

Results and Discussion

Data Analysis

Our results are presented in graphs as found below.

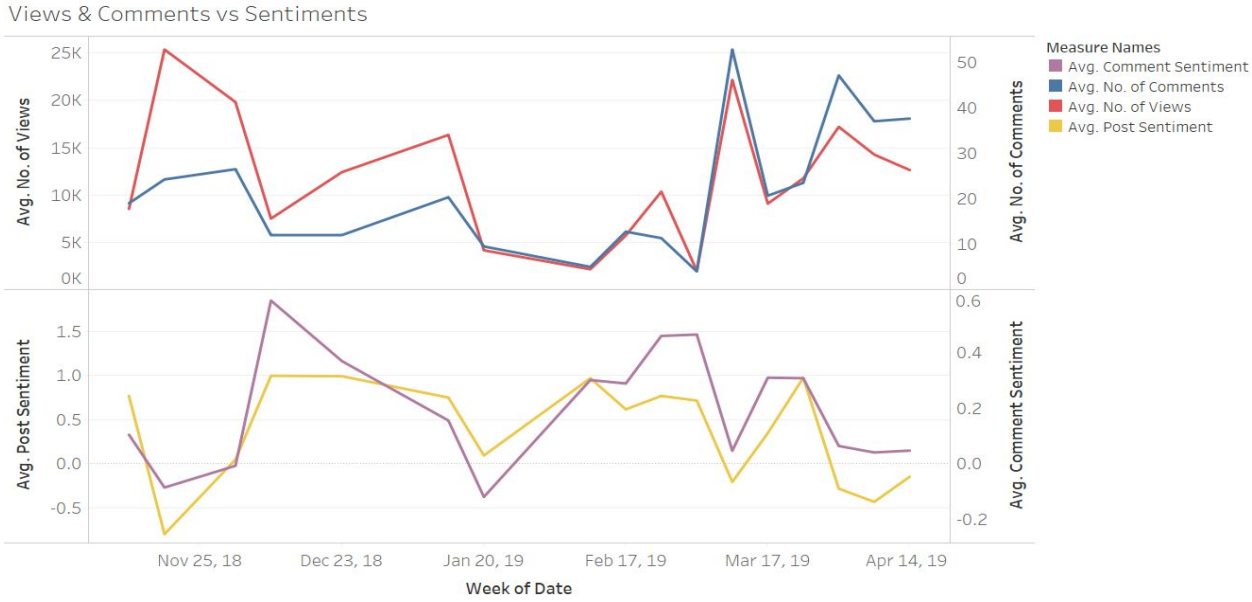

Fig 1.1: Graph of Views, Comments, and Sentiments over Time/Week

We observed that the No. of Views and Comments are directly related, as both graphs have similar shapes when plotted over time (red and blue lines). This agrees with the intuition that when viewership is high, comment rate is also high and vice versa. We observe a similar trend between Post and Comment Sentiments. They are also directly related, with Post and Comment Sentiments increasing and decreasing together (yellow and purple lines). This shows that on a day where Post Sentiment is high, the Comment Sentiment will be high and vice versa. Our most interesting observation is that negative sentiment (Eg. Crime/Bad Behavior) often suggests higher viewership, comment rate and lower comment sentiment. This shows that users are generally more attracted to articles which are with negative sentiment, and comment with similar negativity in sentiment.

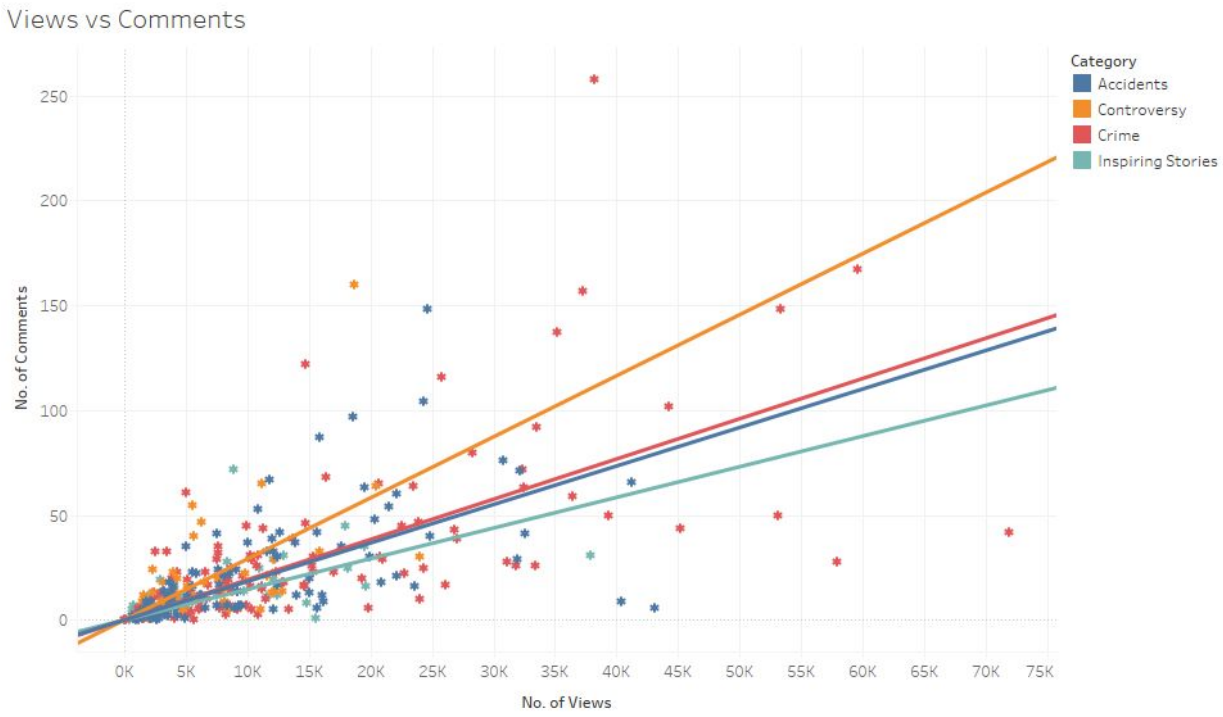

Fig 2.1: Graph of No. of Comments Against No. of Views, broken down by category

$$Gradient = \frac{\Delta No.\ of\ Comments}{\Delta No.\ of\ Views}$$

We plotted the trend lines for each category where y-intercept=0, and observed the gradients of the graphs. Articles in the controversy category have the highest views-to-comment ratio while articles in the inspiring stories category have the lowest views-to-comment ratio. One possible explanation is that Stomp users are more drawn to comment on controversial stories, given how there is more room for discussion.

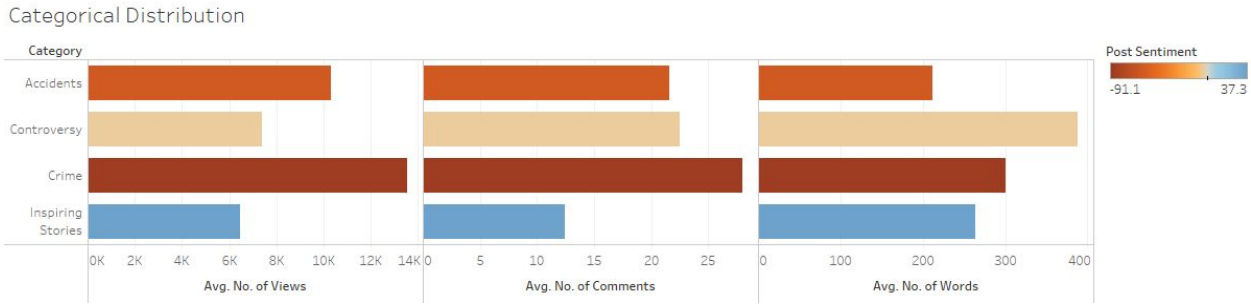

Fig 2.2: Categorical Breakdown of No. of Views, Comments and Words

The bar charts above (Fig 2.2) show the average number of views, comments and words of articles in the four categories. Controversy stories have the most words, while

accident stories have the least. Somewhat surprisingly, crime stories have the most views and comments, even though they contain the second most words behind Controversy stories. Inspiring stories have the fewest views and comments, which agrees with the result in Fig 2.1.

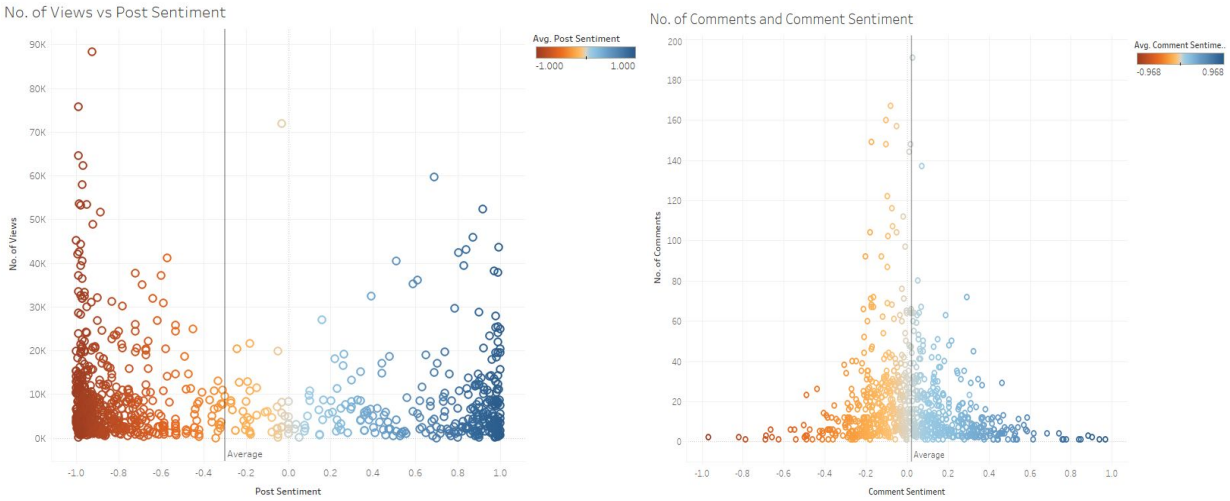

Fig 3.1 and 3.2: No. of Views against Post Sentiment and No. of Comments against Comment Sentiment

Fig 3.1 and 3.2 are scatter plots, where each point represents a single article. The colour represents the sentiment, as shown by the legend. As observed from comparing Graphs of Fig 3.1 and 3.2, Post Sentiment (Fig 3.1) is generally more extreme as compared to Comment Sentiment (Fig 3.2). In Fig 3.1, most articles are congregated at the -1 and 1 sentiment values, and we also observe that these articles also have high viewership. However, in Fig 3.2, more articles have dilute comment sentiment as the articles are congregated toward the centre. From this, we can conclude that user generated sentiment (from comments) are generally less extreme as compared to post sentiments. This implies that most of the negative sentiment on Stomp is contributed by the posts, not comments, contrary to popular belief.

Prediction

After we analysed the data on Stomp, we developed a Predictor using the Naive Bayes Classifier. We trained the classifier on 440 articles and tested it on 110 articles, and the results are as follows.

```
Accuracy = 0.7181818181818181
440 training datasets, 110 testing datasets.
No. of Words = 76542
Most Informative Features
       contains(helping) = True           Inspir : Accide =     26.6 : 1.0
        contains(jailed) = True            Crime : Accide =     21.4 : 1.0
      contains(accident) = True           Accide : Contro =     21.2 : 1.0
  contains(motorcyclist) = True           Accide : Crime  =     20.7 : 1.0
         contains(days) = True            Contro : Accide =     19.4 : 1.0
      contains(junction) = True           Accide : Crime  =     19.0 : 1.0
        contains(guilty) = True            Crime : Accide =     18.4 : 1.0
        contains(review) = True           Contro : Crime  =     18.0 : 1.0
       contains(getting) = True           Contro : Crime  =     18.0 : 1.0
         contains(years) = True            Crime : Accide =     17.8 : 1.0
None
Using Naive Bayes Classifier from NLTK
```

Fig 4.1: Results of Naive Bayes Classifier

From Fig 4.1, we observe that there are certain distinctive words which help us separate the articles into categories. Based on the words in a given text, we calculate the probability of each category and assign the story to the category with the highest probability, using the Naive Bayes Theorem. Although our shown accuracy is 71.8%, the actual accuracy is higher as many categories overlap, and sometimes there is more than one category the article can fall in. Therefore classifier has sufficient accuracy to predict user reactions.

Based on an article, we would generate the Post Sentiment using VaDER Sentiment Analysis. Using the classifier mentioned above, we can obtain the category, hence we are able to suggest a possible comment sentiment based on the average comment sentiment of each category. In addition, we can also use this method to predict the number of views and comments, found from the data analysis earlier.

Conclusion

Using Tableau Data Analytics, we have concluded that Singaporeans prefer articles of lower sentiment. In addition, we found that posts had more extreme sentiments than comments, hence we needed a way to help authors temper their articles.

With our predictor, the author can gauge and predict the user reaction before posting the article, and this algorithm can be applied to hate speech detection and prevention. This is extremely useful and efficient with the huge volume of opinion articles every day.

Further work can be done to include opinion mining or to use other machine learning techniques to increase accuracy. This algorithm can also be expanded to other countries, as this research mainly focuses on Singapore.

With our results, we created a website at https://sites.google.com/view/sentimentanalysisofstomp/ so that authors and social media companies can learn more about social media sentiments and input their articles to predict the user reaction. We hope that this platform can help to increase media objectivity and prevent hate speech.

Bibliography

Below is the list of references we have cited.

1. Holtz, P., Kronberger, N., & Wagner, W. (2012). Analyzing Internet Forums. *Journal of Media Psychology,24*(2), 55-66. doi:10.1027/1864-1105/a000062
2. Low, E. (2014, April 28). Why SPH is keeping STOMP. Retrieved from https://www.marketing-interactive.com/sph-keep-stomp/
3. Noor, R. (2016). Citizen Journalism vs. Mainstream Journalism: A Study on Challenges Posed by Amateurs. *Athens Journal of Mass Media and Communications,3*(1), 55-76. doi:10.30958/ajmmc.3.1.4
4. Pendry, L. F., & Salvatore, J. (2015). Individual and social benefits of online discussion forums. *Computers in Human Behavior,50*, 211-220. doi:10.1016/j.chb.2015.03.067
5. Martinsson, J. (2009). The Role of Media Literacy in the Governance Reform Agenda. *Communication for Governance and Accountability Program (CommGAP),*50300th ser. Retrieved July 21, 2019, from http://documents.worldbank.org/curated/en/891511468331267009/pdf/503000WP0Box341ia0Literacy01PUBLIC1.pdf
6. Sarlan, Aliza & Nadam, Chayanit & Basri, Shuib. (2014). Twitter sentiment analysis. 212-216. 10.1109/ICIMU.2014.7066632
7. Sanaz. (2012). Online Forums Hotspot Prediction Based on Sentiment Analysis. *Journal of Computer Science,8*(8), 1219-1224. doi:10.3844/jcssp.2012.1219.1224
8. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

All graphs were generated by the authors using Tableau.