RESEARCHERID
THOMSON REUTERS

[Chen *et al.*, 5(5): May, 2018] ISSN 2349-0292
Impact Factor 3.802

GLOBAL JOURNAL OF ADVANCED ENGINEERING TECHNOLOGIES AND SCIENCES

# SENTIMENT ANALYSIS OF NEWS ARTICLES AND ITS COMMENTS: A NATURAL LANGUAGE PROCESSING APPLICATION

**Andy W. Chen***

* Sauder School of Business, University of British Columbia, Canada

**DOI: 10.5281/zenodo.1255804**

## ABSTRACT

In this paper I present the use of the VADER (Valence Aware Dictionary and sEntiment Reasoner) model in conducting sentiment analysis of news articles. I explore the sentiments of articles in The Global and Mail and comments made by readers. The dataset contains over 10,000 articles from 2013 to 2016 as well as over 660,000 comments. I find that during this period, the sentiments stay rather consistent and average around a score of 0 (neutral), while the sentiments of the comments follow a bimodal distribution, with around 10% of users consistently giving highly positive comments and another 10% giving highly negative ones.

**KEYWORDS**: Natural Language Processing, Sentiment Analysis, VADER, Machine Learning, Data Science

## INTRODUCTION

A large amount of text is generated each day on social media, news, personal and corporate websites. These data contain valuable information about the authors such as their preferences, sentiments towards certain topics, and demographics. However, the task of uncovering the information manually is impractical. Automated algorithms for extracting information from textual data. In this paper, I explore the use of sentiment analysis, which is a branch of natural language processing, in analyzing new articles. In particular, I conduct sentiment analysis on over 10,000 articles in The Globe and Mail and the comments of these articles. I compare the sentiments over time to uncover any trend and explore the behavior of the authors of the articles and commenters. I find that the average score the of sentiment stays consistent between 2013 and 2016, with a few sudden in increases and decreases. The distribution of the sentiments for articles has a mean of 0 and follows a normal distribution, while the sentiments for the comments has a bimodal distribution.

Related work in this area includes a paper by Ni et al.[1], who develop a structured approach called max-margin structure (MMS) for extraction information from text. Hirshcberg and Manning[2] study the success and challenges of applying natural language processing in machine translation, text mining in social media, and sentiment analysis. Imran et al.[3] study the algorithms for parsing textual data, optimizing work in handling information, and prioritizing types of information in the context of monitoring social media in a mass emergency. Hinrichs et al.[4] propose a deduction-based method for translating between German and English. Middleton et al.[5] propose a platform for monitoring social media during natural disasters and evaluate the model by comparing it to the official impact assessment released by the US National Geospatial Agency.

## METHODS

I use a data set with 10,339 opinion articles published in The Globe and Mail between January 2012 and December 2016. The data set also contains 663,173 comments in 303,665 comment threads. The data is divided into two parts: raw data and annotated data. The raw data includes the original articles, comments, and comment threads. The annotated data includes a score made by human readers to indicate whether a comment is constructive or toxic.

I conduct sentiment analysis on the raw articles and comments using VADER (Valence Aware Dictionary and sEntiment Reasoner), which is based on a dictionary of a set of words with positive or negative sentiment scores. VADER calculates the sentiment score of a document by summing up the sentiment scores of each word in the dictionary.

ISSN 2349-0292
Impact Factor 3.802

## RESULTS AND DISCUSSION

I find that Between 2013 and 2016, the trend of sentiment scores is rather stable between -0.25 and 0.25 (Figure 1). There are some sharp peaks and troughs reach 0.5 and -0.5. I also find that the sentiments of the comments are correlated with the sentiments of the articles. The peaks and troughs in the sentiment scores correspond to some positive or negative articles. We I also find that the sentiments of the articles are average 0 and resembles a normal distribution (Figure 2). The sentiment of comments is also centered around 0, with approximately 10% of commenters consistently making negative comments and 15% consistently making positive ones (Figure 3).

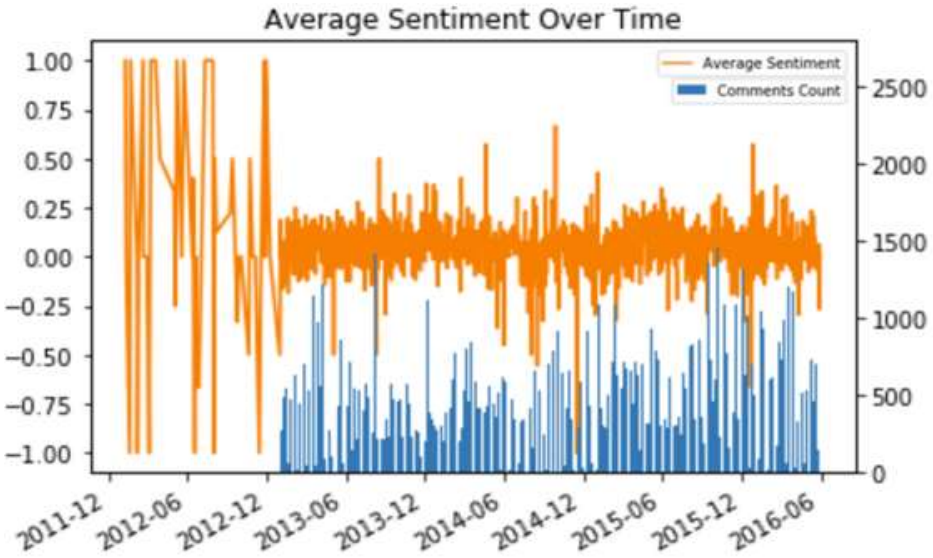


***Figure 1. Average Sentiment over Time***

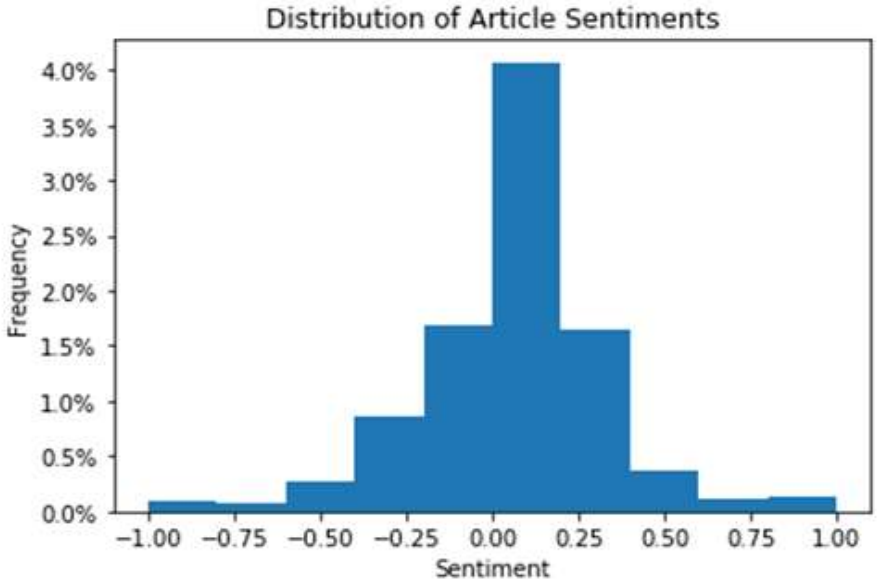


***Figure 2. Distribution of Article Sentiments***

ISSN 2349-0292
Impact Factor 3.802

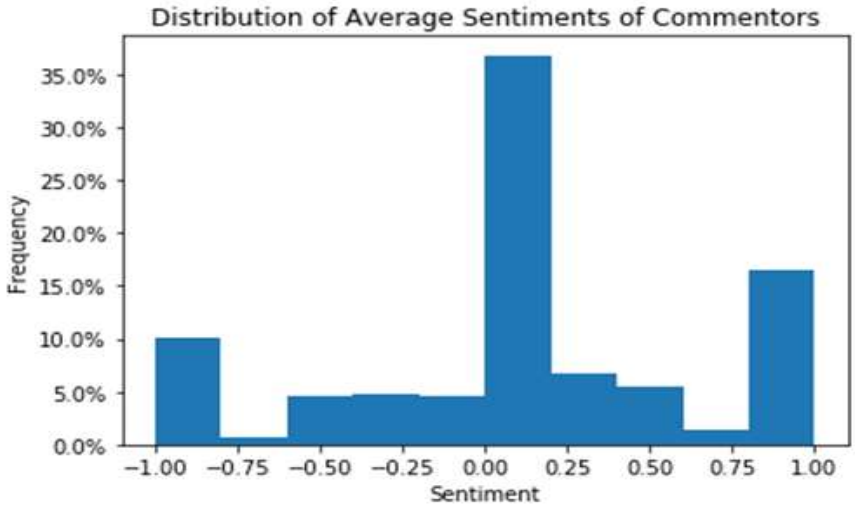


***Figure 3. Distribution of Average Sentiments of Commenters***

The stable pattern of the sentiments over time is reasonable as the media is expected to remain objective, and the number of articles that induce positive and negative sentiments should balance out. The peaks and troughs of sentiments are due to unusual news. For example, on 2014-09-06, the sentiment score reached a peak of 0.034, and the articles released that day contained phrases such as 'I love it'. On 2014-06-01, the sentiments reached a low of -0.45, and the articles contained phrases such as 'misogyny', 'terrorism', and 'abortion'. The normal distribution of article sentiments is expected as reporters are often expected to remain neutral when reporting.

## CONCLUSION

In this paper I describe the sentiment analysis conducted on over 10,000 articles in The Globe and Mail and over 660,000 comments directed at these articles. The results show that the average sentiments of the articles are consistent and centered around 0 over time, while the sentiments of the comments have a bipolar distribution, with 10 to 15% of the commenters consistently giving positive and negative comments. As a future extension, it would be interesting to the find the relationship between sentiments and certain topics or entities in the articles. For example, is there a particular person or organization that elicits consistently positive or negative comments?

## REFERENCES

[1] Ni Y, Saunders C, Szedmak S. The Application of Structured Learning in Natural Language Processing. Machine Translation. 2010;24(2):71-85.
[2] Hirschberg J, Manning CD. Advances in Natural Language Processing. Science. 2015;349(6245):261-266.
[3] Imran M, Castillo C, Diaz F, Vieweg S. Processing Social Media Messages in Mass Emergency: A Survey. ACM Computing Surveys. 2015;47(4):67.
[4] Hinrichs E, Henrich V, Barkey R. Parsing Brief and Informal Messages, Handling Information Overload, and Prioritizing Different Types of Information Found in Messages. Language Resources and Evaluation. 2013;47(3):839-858.
[5] Middleton SE, Middleton L, Modafferi S. Real-Time Crisis Mapping of Natural Disasters Using Social Media. IEEE Intelligent Systems. 2014;29(2):9-17.