Unraveling Public Perceptions of Self-Driving Cars through Social Media Analytics

Abstract

Sentiment analysis and named entity recognition are fundamental components in social media analysis. This paper presents an approach to building a social media analysis pipeline for sentiment analysis and named entity recognition. We collected public's discussions about self-driving cars from Twitter and applied various techniques including rule-based sentiment analysis, transformers, and support vector machines and compared the effect of these different approaches. Additionally, we employed named entity recognition to extract the car manufacturer brands.

1 Introduction

Since the mid-1980s, the development of self-driving cars has been a topic of research and experimentation, with numerous universities, research centres, and companies from the automotive and other industries working to advance this technology(Badue et al., 2021). As technology continues to evolve, it has garnered significant attention and discussion on various online platforms, including social media(Kohl et al., 2018).

Social media applications, such as Twitter, have gained tremendous popularity as platforms for users to express their thoughts and engage in discussions pertaining to current events (Khan et al., 2020). This has created an unprecedented opportunity for researchers to collect and analyze vast amounts of data on social behavior and preferences, which we called social media analysis(SMA) (Fan and Gordon, 2014). SMA is applied across various disciplines, including politics and public opinion research(Anstead and O'Loughlin, 2015), as well as in the business and industry sectors. Doing SMA can help businesses evaluate the competitive environment(He et al., 2013), observe brand communities(Gensler et al., 2013), and provide an effective mechanism for achieving marketing objectives and strategies(Rahardja, 2022).

To obtain a more comprehensive understanding of the brand choices, apprehensions, and overall acceptance of self-driving vehicles by the general population of drivers globally, we proposed using the SMA pipeline to analyse collected Twitter reviews to get insights about the public sentiments towards self-driving cars, the most recently talked about brands and the people's attitudes towards these.

2 Related Work

2.1 Text Preprocessing

In the field of Natural Language Processing (NLP), pre-processing techniques such as stemming, lemmatisation, and removing stopwords are widely used. Stemming and lemmatisation aim to reduce the inflectional forms of words and retrieve their basic meanings. The former refers to a crude heuristic process of cutting off a word's suffix or prefix to retrieve its basic form, while the latter requires morphological analysis (Schütze et al., 2008). Stopwords refer to a set of commonly-used words in a language, which usually have little lexical content (Bird et al., 2009). The removal of stopwords usually helps to retrieve the key information of a text. However, experiments shown that removing stopwords can affect the output of sentiment classification tasks (Saif et al., 2012).

2.2 Named Entity Recognition

Named Entity Recognition (NER) is an essential task in Natural Language Processing (NLP) which involves detecting, locating and categorising important nouns or noun phrases in the text. NER has various applications, including Information Extraction and Question Answering. Currently, there are two major challenges in NER: detecting the boundary of named entities and accurately classifying them (Mohit, 2014). The first challenge requires algorithms to correctly identify the boundaries of named entities, for example, distinguishing between the car manufacturer Ford and the verb ford to cross a river (Zheng et al., 2019). The second challenge involves correctly classifying named entities, such as distinguishing between the scientist Nicola Tesla and the car manufacturer Tesla(Ratinov and Roth, 2009). The approaches to NER mainly include supervised methods, namely machine learning approaches, and unsupervised methods, such as using

pre-defined entity rules (Sharnagat, 2014).

2.3 Sentiment Analysis

Sentiment analysis can be viewed as a branch of computational linguistics, natural language processing, and text mining, as it involves analysing and understanding language at a computational level. Additionally, many of the tasks in Sentiment Analysis can be thought of as classification where text is categorized into sentiment classes such as positive, negative, or neutral (Pang and Lee, 2004). Sentiment classification is used in various applications, including social media monitoring and customer feedback analysis. It can be handled in a document (Turney, 2002), a sentence (Kim and Hovy, 2004) or a phrase level (Agarwal et al., 2009). Term frequency has been a key focus in traditional Information Retrieval systems for a long time, and the widely-used TF-IDF (Term Frequency - Inverse Document Frequency) measure has been used since its introduction by Jones in 1972 (Sparck Jones, 1972). Gilbert's development of VADER (Hutto and Gilbert, 2014) has enhanced the sensitivity of sentiment analysis to expressions in social media contexts and has improved the effectiveness of traditional sentiment lexicons. With advances in technical, machine learning-based methods such as Naive Bayes (Go et al., 2009), Support Vector Machines (Cortes and Vapnik, 1995), and neural networks have been found to be effective for sentiment classification.

Related to this work, where we need to classify the sentiment for each Twitter comment and authors, Go et al.(Go et al., 2009) used Naïve Bayes, MaxEnt and Support Vector Machine (SVM) method on Twitter comments. In their experiment, they found that SVM outperforms other classifiers.

3 Methodology

The methodology in this study is shown in Fig 1, comprising three parts: 1) Data collecting, 2) Data Preprocessing, 3) Sentiment Classification.

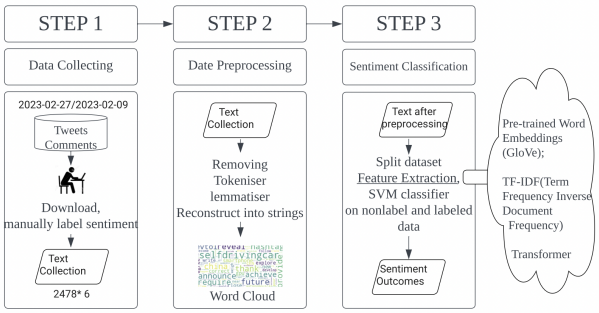

Figure 1: Flow chart for SML pipeline

3.1 Data Collecting

Tweepy (Harmon et al.) is a Python library for accessing and interacting with the Tweeter API. This project used Tweepy to collect the data published from 2023-02-27 to 2023-03-09 based on a list of keywords and hashtags. To retrieve more highly relevant data, in accordance with existing linguistic norms, the list considered both singular and plural hashtags, such as '#selfdrivingcar' and '#selfdrivingcars', as well as keywords with hyphens, like 'self-driving'. Due to the possibility that a single tweet contained two or more keywords or hashtags in the list, the project removed duplicated tweets from the collected data. Finally, there were 2474 distinct data including date, username, location, and the content of the tweets. To explore more sentiment analysis models, we manually labelled whole data collection with 0 representing negative, 1 representing neutral, and 2 representing positive.

3.2 Data Pre-Processing

Our work on pre-processing of raw tweet text used the Natural Language Toolkit (NLTK) library in Python (Bird et al., 2009). The pre-processing step involves removing non-ASCII characters, punctuation, hashtags, URLs, and usernames to obtain clean text. Tokeniser and lemmatiser provided by the NLTK library are then applied, and stopwords are removed. We then tokenise the tokens and reconstruct them into strings in preparation for sentiment analysis. The stopwords we removed are: 'english', 'from', 'subject', 're', 'edu', 'use', 'via', and 'like'. As illustrated before, the removal of stopwords will affect the output of the sentiment analysis, so we manually chose the stopwords. After pre-processing, we obtained the following word cloud, which showed the most-mentioned words in clean text format.

Figure 2: Pre-Processing Result for Data Cleaning

3.3 Sentiment Classification

The SMA pipeline integrates feature extraction, classification and model comparison after the preprocessing stage is completed to provide comprehensive sentiment analysis results. We used three feature extraction

methods in this study: Pre-trained Word Embeddings (GloVe), TF-IDF (Term Frequency-Inverse Document Frequency) and Transformers (attention mechanisms) (Vaswani et al., 2017). Each of these methods extracts valuable information from the dataset differently. The classifier, support vector machine, will then divide the dataset into training and testing sets for each feature extraction method, train an SVM classifier on the training set using the extracted features, and evaluate the performance of the SVM classifier on the testing set by calculating metrics such as accuracy. Finally, we will compare the performance of the SVM classifiers with different feature extraction techniques and the impact of using non-labelled (VADER) and labelled data (Manually).

4 Results and Discussion

4.1 Dominant sentiment towards self-driving cars

Table 1 shows the experimental results for each of the four models using different labelling methods and word embedding methods. We used SVM as the classification component for all models. Firstly, the results of using VADER labelling and manual labelling data with the same word embedding method to pass the same classification model are tested. The results of manual labelling have higher accuracy than VADER labelling. In this case, manually labelling data is used for the next two experiments for testing the performance of different word embedding methods. Finally, it's found that the highest correct rate of 71% was obtained using the manually labelled dataset with TFIDF as feature extraction and SVM as classifier, while the lowest correct rate of 63% was obtained using the Transformer Word Encoding method.

Labelling Method	Model	Accuracy
VADER	TFIDF+SVM	69%
manual	TFIDF+SVM	71%
manual	Pre Train+SVM	66%
manual	Tansformer+SVM	63%

Table 1: Experiment results for four Sentiment Classification Models

Results are analysed below: 1.The results of VADER might be affected by spelling and grammatical errors, thus neglecting important words and resulting in a lower correct rate. 2. The VADER algorithm cannot understand sarcasm or irony, but only the intuitive meaning of each word(Hutto and Gilbert, 2014). In contrast, manual Labelling's approach allows for contextual analysis and the understanding of more flexible linguistic expressions such as parody and irony. However, for our experiments, four people were involved in labelling 2474 comments,indicating four different classification principles, which can lead to artificial misclassification and thus reduce the accuracy of the model. Overall, while machine labelling can be efficient and scalable, manual labeling can provide higher accuracy and a more nuanced understanding of the sentiment expressed in the data.

Lower accuracy getting from Pre Train and Transformer's word embedding methods in this experiment might because of the limitation of data size. For the Pre Train method, the GloVe algorithm is used to embed the dataset. However, due to the 2474 data we have, size glove.6B word vector with domain wikipedia2014+Gigaword is chosen to embed instead of the word vector with the domain of Twitter. Therefore, some words expressing emotions in Twitter cannot be recognized because of domain mismatch, thus reducing the accuracy. For the Transformer methodit usually requires a lager amount of training data for self-attention mechanism to ensure that the relevant patterns and relationships in the data are fully learned(Vaswani et al., 2017). In summary, word embedding using Pre Train and Transformer might performance better if dataset is expanded.

4.2 Sentiment analysis over time

Fig 3 depicts the daily sentiment results processed by VADER to analyse the trend of sentiments over time. All three sentiments showed a dynamic equilibrium. Except for 02-28, the dominant sentiment was always neutral. It can be seen that the neutral sentiment was more stable than the negative and positive sentiments, and decreased slightly over the period. The positive sentiments were lower than neutral sentiments but higher than negative sentiments. Although the positive sentiment vibrated obviously in 02-28 and 03-05, the overall trend is flat. When the percentage of positive sentiment increased significantly, the percentage of negative sentiment decreased significantly. The negative sentiment trended slightly upward in general.

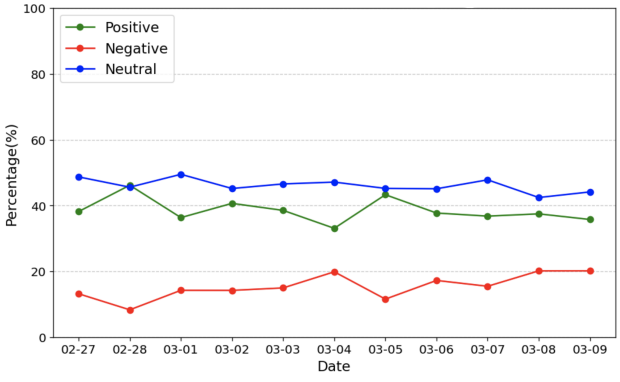

Figure 3: Sentiments Analysis over Time

Considering the data and Fig 2, one possible reason for the vibration of the percentages of positive and negative sentiments in 02-28 and 03-05 could be the advertisements. In the advertisements, there are many positive words, which may influence Vader and result in 'positive' classification results. Because the project does not filter our the tweets of advertisements, if the advertisers advertise regularly, there are possibly obvious vibrations of the sentiments in a specific period.

Therefore, in order to get more valuable sentiments or public opinions for analysis, the project can perform additional data cleaning, such as finding approaches to detect and remove advertisements from data. On the other hand, as the data collection was only from 2023-02-27 to 2023-03-09, the sentiments generally trended to be flat. To get more objective and valuable sentient analysis results, the project should collect data over a longer time span, such as 2022-01-01 to 2023-01-01, and then observe the sentiment trends.

4.3 Recently talked brands

Fig 4 shows the average compound of each brand based on the results of sentiment analysis and named entity recognition. To obtain more reliable analysis results, Fig 5 depicts how many times the named entity of a brand is mentioned in the collected data. According to Fig 5, Tesla is the most mentioned brand.

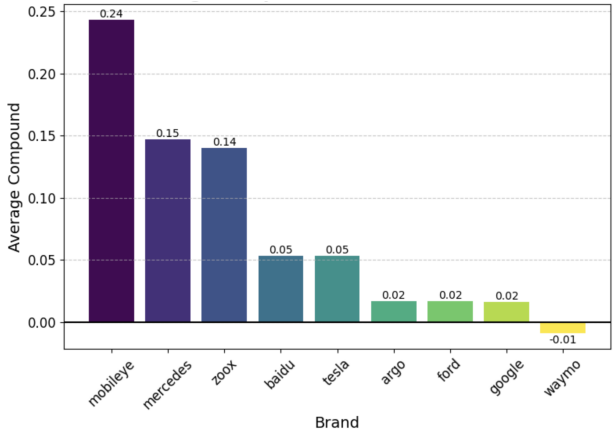

Figure 4: Average Compound Sentiment for Brands

Based solely on the information in Fig 4, Mobileye is the most trusted brand. However, according to Fig 5, the occurrences of brands are highly imbalanced, and Mobileye has only 3 occurrence. Taking Baidu and Tesla for examples, despite having the same average compound of 0.05, Tesla has 112 occurrences while Baidu has only 1. One possible explanation is that Baidu is a Chinese brand, whereas Twitter is prohibited in China. By checking the tweet mentioned Baidu from the source data, it can be seen that the author of the tweet is seeking to promote the progress of China's technological development. Despite mentioning Baidu with a positive sentiment, it is believed that the reference value of sentiment of this tweet is very low and may contain bias due to the low data volume. Tesla's results, on the other hand, are more reliable and valuable for analysis because they are supported by a large amount of data.

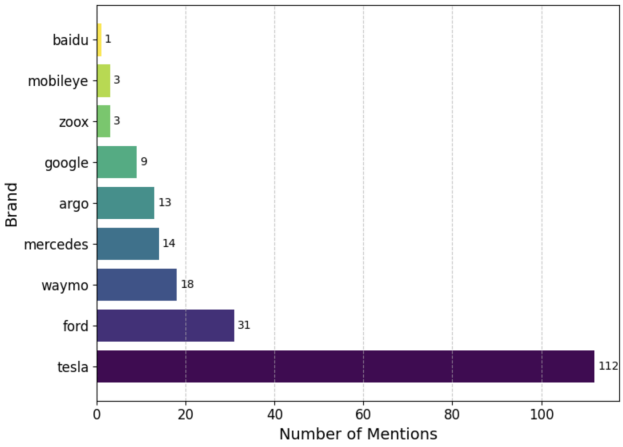

Figure 5: Occurrence of Bands in Tweets

In summary, people tended to trust Mercedes, with occurrence of 14 and average compound of 0.15, as well as Tesla, with occurrence of 112 and average compound of 0.05, during the period we experimented.

5 Conclusion

In this project, a pipeline for social media analysis (SMA) of people's opinions on self-driving cars was built. The pipeline includes pre-processing techniques, such as lemmatisation and stopwords removing from the raw tweet text. Various approaches, like VADER, Transformers and SVM, were employed to conduct the sentiment analysis. Additionally, the named entity recognition provided by SpaCy was used to identify the most discussed and trusted brands. Overall, the sentiment towards self-driving cars was found to be neutral, with slightly more people holding a positive view than a negative view. However, our work also encountered some limitations, such as the small size of the dataset and the use of incompatible domains of word embedding weights, which may lead to a biased result and poor performance of the trained model. To improve the results, retrieving more data from older tweets can be considered as a potential measure. Furthermore, adverts may be considered as a potential new category for sentiment analysis in the future, given that a significant proportion of them exhibit positive sentiment.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 24–32.

Nick Anstead and Ben O'Loughlin. 2015. Social media analysis and public opinion: The 2010 uk general election. *Journal of computer-mediated communication*, 20(2):204–220.

Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M Paixao, Filipe Mutz, et al. 2021. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Weiguo Fan and Michael D Gordon. 2014. The power of social media analytics. *Communications of the ACM*, 57(6):74–81.

Sonja Gensler, Franziska Völckner, Yuping Liu-Thompkins, and Caroline Wiertz. 2013. Managing brands in the social media environment. *Journal of interactive marketing*, 27(4):242–256.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Harmon, Joshua Roesslein, and other contributors. Tweepy.

Wu He, Shenghua Zha, and Ling Li. 2013. Social media competitive analysis and text mining: A case study in the pizza industry. *International journal of information management*, 33(3):464–472.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Rijwan Khan, Piyush Shrivastava, Aashna Kapoor, Aditi Tiwari, and Abhyudaya Mittal. 2020. Social media analysis with ai: sentiment analysis techniques for the analysis of twitter covid-19 data. *J. Crit. Rev*, 7(9):2761–2774.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

Christopher Kohl, Marlene Knigge, Galina Baader, Markus Böhm, and Helmut Krcmar. 2018. Anticipating acceptance of emerging technologies using twitter: the case of self-driving cars. *Journal of Business Economics*, 88:617–642.

Behrang Mohit. 2014. *Named Entity Recognition*, pages 221–245. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Untung Rahardja. 2022. Social media analysis as a marketing strategy in online marketing business. *Startupreneur Bisnis Digital (SABDA Journal)*, 1(2):176–182.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155.

Hassan Saif, Yulan He, and Harith Alani. 2012. Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11*, pages 508–524. Springer.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Rahul Sharnagat. 2014. Named entity recognition: A literature survey. *Center For Indian Language Technology*, pages 1–27.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Peter D Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changmeng Zheng, Yi Cai, Jingyun Xu, HF Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.