ISSN 2007-9737

Topic Modelling and Sentiment Analysis via News Headlines, NLP Methods on Australian Broadcasting Commission

Zaur Gouliev , Fernando Perez-Tellez*

Technological University Dublin,
School of Enterprise Computing, Digital and Data, Dublin,
Ireland

x00205702@mytudublin.ie, fernando.pereztellez@tudublin.ie

Abstract. The main aim of this paper is to provide a holistic overview, implementation and comparison of some of the main supervised and unsupervised machine learning methods that are used in natural language processing for extracting topics and sentiment from headlines. This paper employs supervised learning methods such as logistic regression, supper vector machine classifier (SVM) and unsupervised learning methods such as K-means clustering and Latent Dirichlet allocation (LDA). To demonstrate these NLP applications, an extensive dataset of one million news headlines is used provided online by the Australian Broadcasting Commission which contains 17 years of news headlines, which provides for rich analysis. Our results show that logistic regression based models which use lexicon-based emotion classifiers score very highly in accuracy for sentiment analysis, reaching 93% and clustering-based techniques K-means scored 75% for topic modelling. An detailed explanation of these methods, along with limitations, assumptions, ethical considerations and suggestions of future work are discussed.

Keywords. News headlines, machine learning, natural language processing, sentiment analysis.

1 Introduction

In our current digital era, the consumption of news has reached unprecedented levels, primarily attributed to the ease of access and the burgeoning volume of news content. A study by Pew Research Center highlighted that nearly half (48%) of U.S. adults frequently rely on social media for news [20].

Similar research in the context of Europe, found that for 2021, approximately 72% of internet users aged 16-74 read online news sites, newspapers or news magazines [5]. Figure 1 shows the percentage increase from 2016 to 2021. The trends from both these research studies suggest this is only but increasing.

This scenario poses a critical question: Given the vast array of news sources available, including social media, websites, and television broadcasts, how can we systematically analyze and comprehend the vast content at our disposal? Machine learning (ML), particularly within the domain of natural language processing (NLP), offers a promising solution to this challenge.

ML techniques are adept at processing and categorizing textual data, demonstrating already a significant potential in sectors like law, healthcare, and finance for organizing text into thematic clusters [9]. This paper proposes utilizing these techniques for the automatic classification of news articles by sentiment and topic, thereby facilitating a deeper understanding of the content we consume.

This paper, in large, is a demonstration of those techniques and analysis of the results. The objective is to explore and identify the most effective ML models, both supervised and unsupervised, for news analysis, aiming to improve media interpretation and consumption. Sentiment analysis within this context aims to discern the emotional tone of news headlines, categorizing

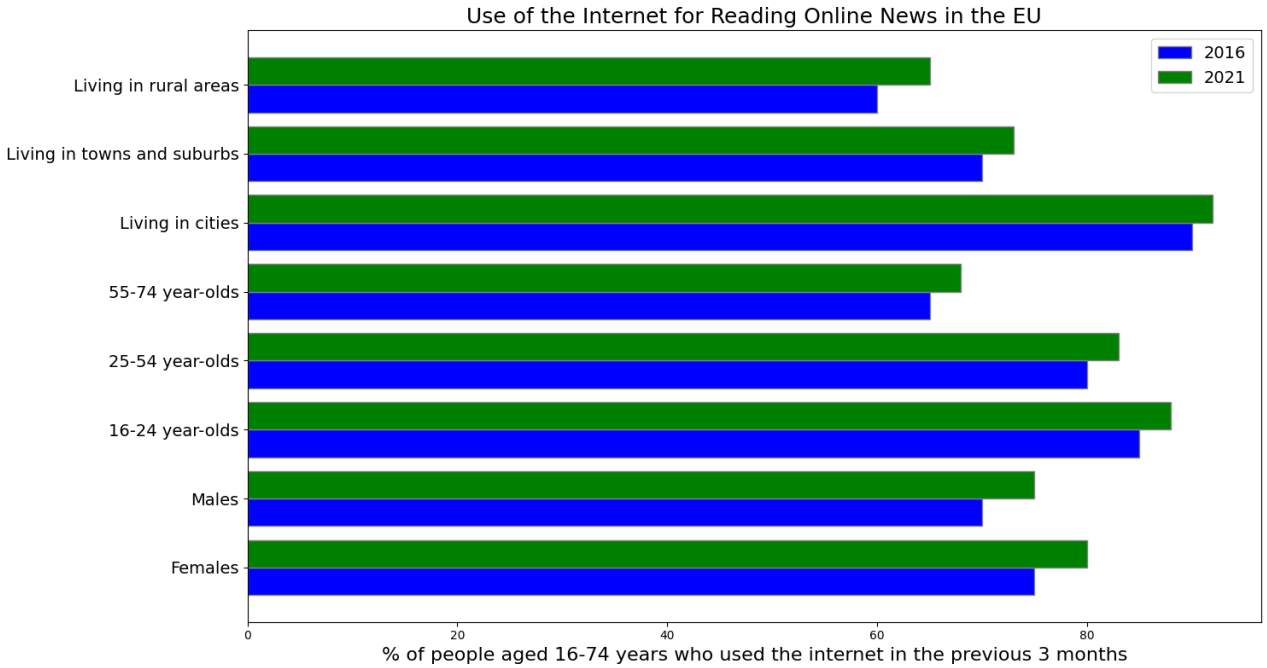

Fig. 1. Use of the internet for reading online in the EU, showing increase from 2016 to 2021 [5]

them as positive, negative, or neutral. This approach provides insights into the emotional undertones of news content, with negative headlines often reflecting anger and positive ones conveying happiness. Topic modeling, on the other hand, identifies the principal themes discussed in news articles. For instance, articles on ceasefire agreements in the Ukrainian-Russian conflict would be categorized under ”international affairs” or ”politics,” while coverage of a Pakistan vs. India cricket match would fall under ”sports.”

However, headlines that blend themes, such as ”Sports cricket star Imran Khan’s controversial remarks on the Kashmir conflict at Kings College” present more complex cases for classification.

The application of ML in this manner allows for an in-depth analysis of public opinions and media attitudes, enabling data scientists to track evolving trends over time through topic modeling techniques. This paper also addresses potential limitations and challenges inherent in applying ML methods to large datasets and discusses the implications of these techniques for enhancing our understanding of the media landscape.

The digital age has not only transformed how news is consumed but also how it’s analyzed and understood. The proliferation of online news sources and social media platforms has led to an information overload, necessitating advanced methods for filtering and comprehending vast datasets. The diversity and volume of news content available online present unique challenges and opportunities for automated content analysis [6].

This paper’s focus on sentiment analysis and topic modeling of news headlines through NLP methods seeks to contribute to this field of study. Sentiment analysis, involves determining the affective state conveyed by a text, which in the context of news, can offer insights into the media’s framing of events and issues [13].

Topic modeling through Latent Dirichlet Allocation (LDA), provides a method for discovering the underlying themes in a large corpus of text (Blei, 2012). By applying these techniques to a comprehensive dataset of news headlines from the Australian Broadcasting Commission, this paper aims to shed light on the predominant sentiments and topics covered over a 17-year period.

Table 1. Accuracy results of studies using supervised learning methods for NLP

Study	Method	Accuracy
Pang et al. (2008) [18]	Logistic Regression	85%
Felciah et al. (2016) [21]	Logistic Regression	88%
Sharma (2020) [22]	Logistic Regression	85%
Muhammad et al. (2022) [16]	SVM	80%
Singh et al. (2018) [23]	SVM	83%
Chaganti et al. (2014) [3]	SVM	93%

Table 2. Accuracy results of studies using unsupervised learning methods for NLP

Study	Method	Metric	Score
Alharbi et al. (2021) [1]	K-means Clustering	Accuracy	87%
Pang and Lee (2002) [19]	K-means Clustering	Accuracy	83%
Ali et al. (2022) [2]	LDA	Accuracy	70%
Valenti et al. (2017) [24]	LDA	Accuracy	76%
Kirill et al. (2020) [10]	LDA	ROC AUC	73%

One of the key challenged of accurately categorizing news content is the nuances of language and the context-dependent nature of sentiment and themes. News headlines, often crafted to be eye-catching and impactful, may convey multiple sentiments or cross several topics.

The mixed nature of such headlines underscores the complexity of the task at hand and highlights the need for sophisticated ML models that can navigate these subtleties.

There is also an ethical dilemma, in so far that ethical there exists biases in news headlines and biases which are inherent in machine learning models so these need to be considered with careful consideration, the biases present in training data can lead to skewed analyses, affecting the accuracy and fairness of conclusions drawn from NLP applications [12]. This paper acknowledges these challenges, aiming to critically assess the performance of various ML models in news analysis, including supervised methods like logistic regression and SVM classifiers, and unsupervised methods like K-means clustering and LDA and conclude with comments on ethical considerations and biases.

2 Literature

The field of Natural Language Processing (NLP) intersects with an array of disciplines, each benefiting from its advancements in unique ways. NLP's ability to parse, understand, and categorize text has made it indispensable in the age of digital information, where data is plentiful but insights are scarce.

The exponential increase in textual data generation [17], underscores the growing relevance of NLP technologies in deciphering the vast amounts of unstructured data produced daily. There are numerous examples of its applications, for example, NLP can improve customer service interactions in the finance sector by analyzing sentiment and categorizing inquiries highly accurately [4].

Similarly, machine learning techniques have been applied in healthcare to categorize and analyze clinical notes for better patient outcomes [8, 7]. In the financial sector, the application of NLP for sentiment analysis illustrates the critical role of understanding market sentiment in making informed investment decisions.

Research [14] found using sentiment analysis to parse through vast quantities of financial news to extract whether bullish (positive) or bearish (negative), provided investors with an overview of the market climate. Such applications underscore NLP's ability to not only manage large datasets but also to extract nuanced insights that are not immediately apparent.

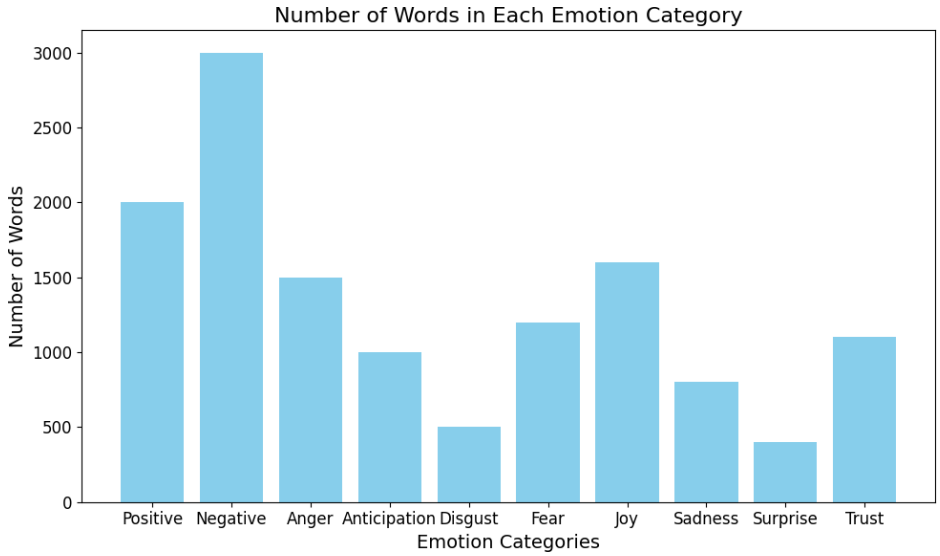

Fig. 2. The number of words in each emotion category

Fig. 3. The words in positive and negative category

2.1 Supervised Techniques

Logistic regression is a commonly used algorithm in machine learning, it is used in many fields for a variety of tasks, things like spam detection are a good example. The reason this is a good example is because an email can either be spam or not spam, and this refers to a binary-based problem and logistic regression is well tuned for solving these binary classification problems where sentiment analysis is one, i.e., negative or positive news. The algorithm is fairly self-explanatory, it works on estimation, based on the probability of an event occurring by mapping the input variables to a continuous output between 0 and 1. These models are trained using labelled data (hence why it is supervised), and they are well known for their efficiency in handling larger datasets with high-dimensional features. Studies previously done using logistic regression in NLP looked at classifying movie reviews as either positive or negative [18].

Fig. 4. A word cloud of common words found in our news headlines

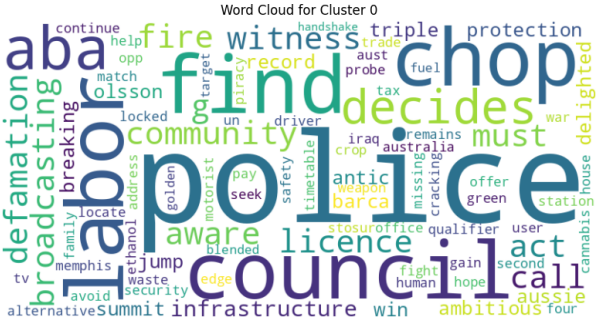

Fig. 5. Cluster 1: Broadcasting and legal matters

Fig. 6. Cluster 2: Law enforcement and safety

These researchers used logistic regression to get the sentiment of a review on a movie. The paper was highly accurate, and the model achieved 85% accuracy on determining the sentiment. Another study used logistic regression to classify online reviews of smartphone products as either positive or negative [21].

The input data was related to negative words or words that show emotion like 'upset', 'disappointed', 'unsatisfied' and they used this to achieve huge accuracy of predicting the sentiment in the reviews (88%).

Due to these high accuracy levels, we will be using logistic regression in our paper to test this. Another more recent paper [22] looked at Amazon product reviews, and accurately predicted the sentiment of the opinion using logistic regression, which also proved to 85% accurate.

Support vector machines (SVM) is another commonly used class of supervised learning methods, it is well known for its classification tasks, it is similar to our previous model, it is a binary classification based tool which solves multiple binary problems, i.e., more than two classes in our input data. It works by finding the most optimal hyperplane that separates different classes in the dataset by finding the maximum margin between the closest data points of each class.

A study done on social media used SVM to classify sentences into positive, negative or neutral categories of sentiment found that SVM had an accuracy of 0.83% [23], which makes it almost as good as our previous discussed logistic regression. The following table shows a summary of results from studies that have used logistic regression and SVM for sentiment analysis and topic modelling. Table 1 shows a summarised version of these studies and their results.

2.2 Unsupervised Techniques

K-means clustering is a very popular and common model that is used in unsupervised machine learning. The algorithm works by grouping the most similar data points in a dataset into clusters (often called k-clusters), the algorithm attempts to minimize the sum of squared distances between data points and their assigned clusters, often called the centroid or centroids.

This method is used widely in fields of data mining and document retrieval. It is useful for grouping similarities in words and sentences which is why it proves so useful. This algorithm is also able to be used in large datasets and high-dimensional features.

A study done by [1] used k-means clustering to identify similar Arabic news articles, they used an approach called bag-of-words to represent the text data and the applied k-means to identify clusters that were similar, they compared these to other

Table 3. Methodology carried out in this research work

Steps	Description of Steps
Importing Packages	Essential libraries like pandas, numpy, and nltk are imported to facilitate data processing and analysis tasks.
Text Preprocessing	Data is cleaned by removing punctuation, numbers, and stop words. Additionally, words are lemmatized to their base form.
Vectorization	Text is converted into numerical form using TF-IDF vectorization, a crucial step to prepare for machine learning algorithms as ML models recognise numbers/vectors.
Clustering	Similar data points are grouped together using K-means clustering, enabling the exploration of common themes within the dataset.
Topic Modeling	Abstract topics within the dataset are discovered using the LDA model, with visualizations aiding interpretation.
Train-Test Split	Data is divided into training and testing sets to evaluate model performance accurately.
Emotion Analysis	Sentiments expressed in text are identified using the NRC Emotion Lexicon, providing insights into emotional content. It is trained using the NRC lexicon.
Logistic Regression	A model is trained to predict sentiment based on text features, leveraging logistic regression for binary classification tasks.
Prediction	Using the trained model, sentiment is predicted for new text data, enabling automated sentiment analysis.

clustering algorithms but found this to be the best performing out of them all, scoring a 87% accuracy. This is also our reason for implementing this algorithm in our paper. Other researchers noted a high accuracy when using k-means for movie reviews [19].

Latent Dirichlet allocation (LDA) is another unsupervised learning method commonly used in NLP for topic modeling. It works very well for identifying topics in a large dictionary of text called a corpus or corpora. It is an unsupervised learning algorithm used very commonly in identifying themes and underlying topics within textual data and comparing them amongst different texts, the way it works is by assuming that each document or text in a given corpus is made of a mixture of topics, and that each topic has a probability distribution over words, the algorithm then identifies the topics from the text data and the probability distribution of each topic in each document. For this reason, it used a lot in identifying hidden patterns in structures of corpus in text data.

One study that used LDA to identify topics in news articles, the researchers applied an LDA algorithm to identify tourism reviews about Marrakech city from TripAdvisor reviews found that the results were highly accurate and found that LDA was effective in identifying meaningful topics within the large corpus of text data, with accuracy of 70% in this case [2].

Table 4. Cluster topics

Cluster	Topic
0	Legal Proceedings and Arrests
1	Community Broadcasting
2	Police Investigations
3	Legal Proceedings and Arrests
4	Government Policies and Decisions

Another paper [24] LDA was used for topic modelling to infer the emotional state of people living with Parkinson's disease, saw results of 76%, this seemed to follow similar results with work by Kirill et al., [10] who looked at how to identify propaganda online using topic modelling, and LDA was also used in this case, and results were predictive, reaching an ROC AUC (Area under the ROC Curve) of 73% .

The ROC AUC curve tells us how much the model is capable of distinguishing between classes, and so the result is strong. Table 2 shows a summarised version of these studies and their results. Each model that is proposed here has its own disadvantages, as well as its own advantages, and much of it is has to do with the quality of data, the ability of the algorithm to determine what is being said, and finally the optimisation of the models.

It is important to emphasis that the field of NLP is relatively new and models are still being created and deployed, thus no model can be perfect as each one has its own performance and are influenced by many factors during the machine learning application process, however some are better than others and this paper attempts to apply this test to a large dataset. There is a growing interest in hyperparameter tuning and fine tuning models to suit specific types of data, in our paper we do not investigate this topic.

3 Data and Methods

In this study, we leverage the Australian Broadcasting Commission: A Million News Headlines dataset [11], spanning a comprehensive period of 19 years, from mid 2003 to the end of 2021.

For consistency, our dataset ignores the first year 2003 and the final year 2021 to incorporate a full year news cycle. The data comprises of around one million news headlines, this dataset is publicly accessible through various platforms, including Kaggle. It is a labelled dataset, which we expect to perform well with supervised learning methods. It contains two key attributes which is the publishing date and the headline text which we will be using in our analysis. A full breakdown of each of the steps taken can be found on on table 3.

In our paper, we will be incorporating the NRC Emotion Lexicon, a comprehensive compilation of English words annotated with associations to eight primary emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), along with two sentiments (negative and positive). The National Research Council (NRC) Emotion Lexicon [15], is combined with the logistic regression model to analyze each headline to determine the sentiment.

This emotion lexicon is widely used database that contains a list of words that are associated with emotions. It contains about 14,000 words that are classified into various emotions, for example, words such as 'enjoying' and 'awesome' will be associated with emotion of joy and happy, while words such as 'shouting' and 'raging' will be associated with angry and sadness which will be matched against our news headlines.

Figure 2 shows the emotion categories, along with number of words, while Figure 3 shows the actual words contained in the positive and negative emotion lexicon, it is worth noting that much of these words are contained in news headlines, noted during our exploratory data analysis.

Our dataset features two primary components which is a publish date and a headline text. The publish date denotes the timestamp indicating the publication date and time of the news articles, while the headline text attribute encapsulates the actual content of the news headlines being disseminated.

4 Results and Evaluation

Based on the experiments that we did on our "Australian Broadcasting Commission: A Million News Headlines" dataset, we find that our logistic model works out the best with an accuracy of

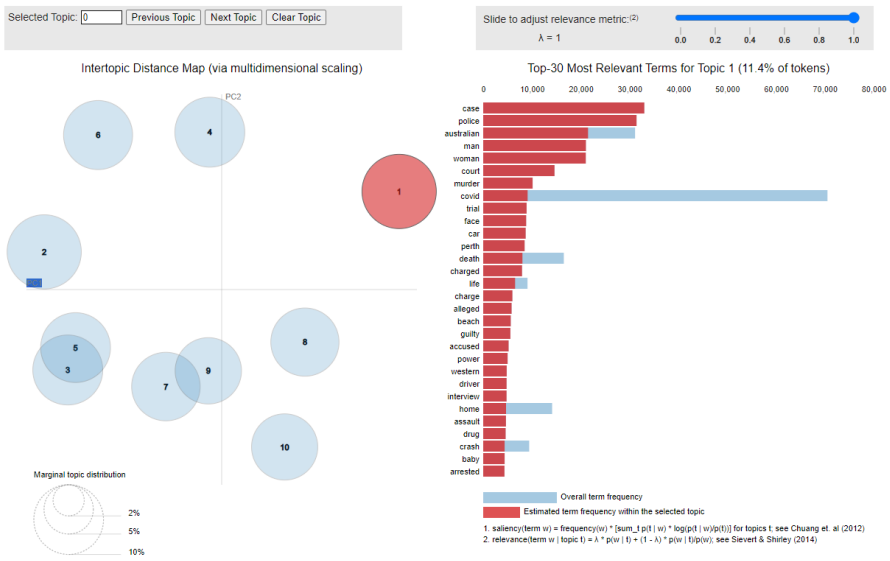

Fig. 7. Topic 1: Emergency services and pandemic news

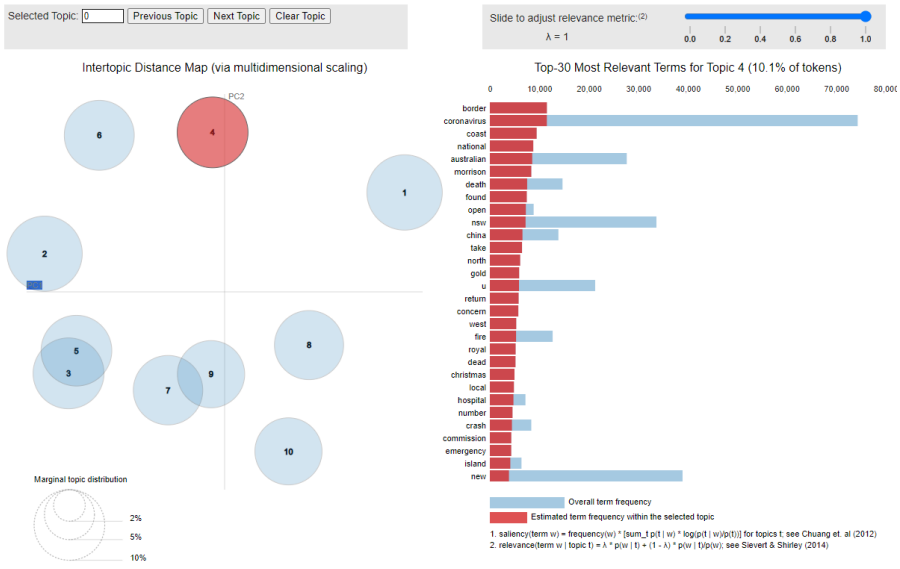

Fig. 8. Topic 2: Policing and safety news

93% on our dataset when combined with our NRC lexicon. We also find the k-means clustering model has an high degree of accuracy at clustering news into different relevant clusters based on a topic, similarly to the LDA model. A list of tables of the topics modelled as well as the sentiment performance is given illustratively below.

4.1 Unsupervised Methods

We had moderately accurate results when using K-Means Clustering, Figure 4 shows us the a word cloud of top words in our dataset of headlines, while figure 5 appears to contain headlines related to legal decisions, community matters,

Table 5. Top 10 topics from our LDA model

Topic	Keywords
1	State and Community Affairs: wales, western, minister, indigenous
2	Events and Incidents: trump, woman, melbourne, bushfire
3	Legal Proceedings and Social Issues: case, man, change, court
4	Economic and Environmental Concerns: home, finance, coast, market
5	Health and Disasters: victoria, health, crash, tasmania
6	Political and International Affairs: sydney, vaccine, election, minister
7	COVID-19 and International Relations: covid, coronavirus, china, border
8	Government and Public Safety: government, restriction, scott, adelaide
9	National Issues and Investigations: australia, australian, death, pandemic
10	Social and Political Movements: queensland, open, national, worker

and broadcasting licensing. Terms like ”decides”, ”community”, ”broadcasting”, and ”licence” suggest legal or administrative proceedings related to community broadcasting. The presence of words like ”fire”, ”witness”, and ”defamation” might indicate legal cases or incidents involving fire and witness testimonies and Figure 6 seems to involve headlines related to law enforcement, safety, and alternative solutions.

Terms like ”police”, ”station”, ”cracking”, ”driver”, and ”safety” suggest law enforcement activities and efforts to ensure public safety. The term ”aboriginal” indicates a focus on issues or events involving indigenous communities, possibly related to law enforcement actions or initiatives aimed at supporting these communities.

Table 4 shows us these clusters. The LDA model performed moderately well at topic modelling, the results and corresponding keywords were somewhat relevant with moderate accuracy but some words were flagged, and we are unsure of why this is.

One possible issue that we might have here is that because we are only analysing news headlines, it does not take into consideration the actual content of the article. This means we would need more detailed information, as that could help us identify nuances and subtleties that might be missed by analyzing only headlines, some headlines could be ironic, or sarcastic and this would mean our algorithm is misunderstanding these. Figure 7 show the top 30 most relevant terms for Topic 1 - 11.4% of tokens and flagged words such as border, coronavirus, coast, national, death, fire, hospital, crash, emergency.

These words could be considered as news topics based on emergency, i.e., the COVID-19 pandemic or accidents that make national news, while Figure 8 shows the top 30 most relevant terms for Topic 4 - 10.1% of tokens, and flagged words such as case, police, man, woman, court, murder, trial, charged, life, charge, alleged, guilty, accused, assault, drug, crash, baby, arrested.

These words suggest that this topic is predominately around news of crime and policing, e.g., murders or court news that make national news. Table 5 gives us a list of all the topics modelled from our LDA model.

4.2 Supervised Methods

Our SVM model performed moderately well for sentiment analysis determining with accuracy of 75%. Our best performing model was the logistic regression with the NRC classifier for sentiment analysis, it reached an accuracy of 93%.

Table 6. SVM model evaluation metrics

Metric	Value
Accuracy	0.7588
Precision	0.7955
Recall	0.7548
F1-score	0.7747

Table 7. SVM classification report

Class	Precision	Recall	F1-Score	Support
0	0.77	0.78	0.78	7081
1	0.70	0.75	0.77	1433
Accuracy			0.76	8514
Macro Avg	0.73	0.72	0.73	8514
Weighted Avg	0.76	0.76	0.76	8514

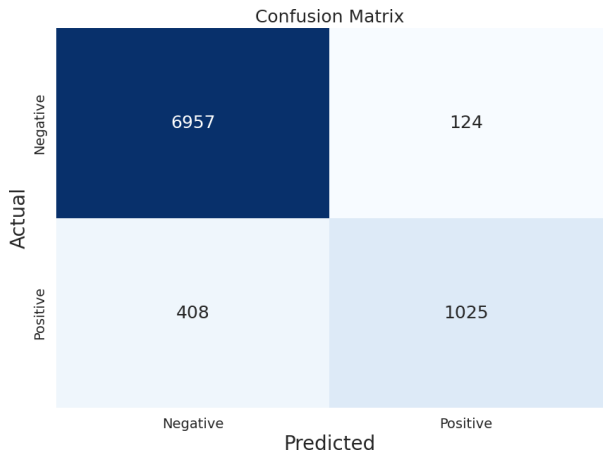

Fig. 9. Confusion matrix of linear regression model with NRC classifier

Table 8. Model evaluation metrics for linear regression model with NRC classifier

Metric	Value
Accuracy	0.9375

We also present an evaluation of this model with mock news headlines to see its performance. The results are presented in the table 8, classification report in table 9 and a confusion matrix presented in figure 9.

5 Risks and Ethical Consideration

In this paper there are risk and ethical considerations, the first of these is to understand that as with any machine learning technique, we need to understand they are not perfect and can result in mistakes both from a performance standpoint (overfitting, underfitting, multicollinearity) as well as from a latent standpoint, where there may be something that is going on that we are unaware of, such as the machine learning algorithm is picking up and we aren't, or the result is happening for reasons why they are associated are not obvious to us but obvious to the model, this is a common problem when modelling emotions and psychometric based models.

In employing logistic regression models alongside lexicon-based emotion classifiers for sentiment analysis, these approaches heavily lean on predefined dictionaries, missing the subtleties and context-dependent variations in emotional expression. This can result in misclassification or oversimplification, especially in intricate or ambiguous news headlines.

These methods often assume that emotions can be neatly categorized, disregarding the multifaceted nature of human emotions. Consequently, the model could struggle to capture nuanced emotional nuances, leading to biased or incomplete analyses.

Beyond modelling risks, we also should note that it is important to remember the ethics of privacy, since we are dealing with news data and often this data contains personal information about individuals, locations, and things that could be used to identify a person or entity.

The key goal is to ensure that the information does not lead to breaking of GDPR type policies, as it is pertinent that privacy is not compromised, one way to do this is to manually inspect the data and another more advanced method is to use data-privacy techniques to mask any potential data that could be carrying personal information.

Due to the vast size of our dataset we could not verify it entirely, and names of politicians, celebrities or people could be mentioned in the news analysis that spans over 17 years, which

Table 9. Classification report for linear regression model with NRC classifier

Class	Precision	Recall	F1-Score	Support
0	0.94	0.98	0.96	7081
1	0.89	0.72	0.79	1433
Accuracy			0.94	8514
Macro Avg	0.92	0.85	0.88	8514
Weighted Avg	0.94	0.94	0.93	8514

Table 10. Sentiment analysis of news headlines on a test headlines data

Headline	Sentiment
Ireland celebrates record-breaking tourist numbers	Positive
Stock market experiences sharp decline amid economic uncertainties	Negative
New study suggests benefits of regular exercise for mental health	Positive
Government announces new measures to tackle unemployment rates	Positive

spans into the data ownership and consent, news organisations have legal ownership of the data they produce and collect, and if we use this data for analysis, we should consider that there needs to be informed consent from each party, since the original news headline creator did not consent to the headline being used for large-scale analysis.

6 Conclusion

This paper reached its aim in trying to showcase how natural language processing techniques can be used to gain useful insight into textual data, in our paper we focused on news headlines, but this can apply to any sort of text data. Our logistic model's ability to be able to classify the sentiment with accuracy of 93% in figure 8, shows how useful this algorithm is, compared to SVMs accuracy at 75% as shown on figure 6. The ability for the k-means clustering model and LDA model to group headlines also makes us consider how it can be applied to more domains than just news headlines.

One plausible reason for the logistic model's superior performance, achieving an impressive 93% accuracy, could be its robustness in handling the binary classification of sentiment.

Logistic regression benefits from a probabilistic approach, which allows it to model the probability of class memberships. Given that sentiment is often a direct reflection of probability—whether a piece of text conveys a positive or negative sentiment—logistic regression's methodology aligns well with the inherent nature of the sentiment analysis task.

The logistic model may have been advantaged by the dataset's characteristics, which appears to be well-suited for a binary logistic approach. The presence of clear, emotive lexicon within news headlines could lead to strong indicators of sentiment, allowing logistic regression to effectively capitalize on these features.

Additionally, the model might be less prone to overfitting when compared to more complex models, especially if the feature space is well regularized, leading to better generalization on unseen data. In contrast, SVM's performance might have been hindered due to its sensitivity to the choice of kernel and regularization parameters.

SVM models can also struggle with large feature spaces, as might be the case with the rich and diverse dataset provided by the Australian Broadcasting Commission, leading to its lower accuracy of 73%. Future research could include utilising the full article and comparing the sentiment of the article to the news headline to see if it matches or if indeed the topic is related, and with more advanced modelling such as deep learning (RNNs, CNNs, LLMs).

References

1. **Alharbi, A. R., Hijji, M., Aljaedi, A. (2021).** Enhancing topic clustering for arabic security news based on k-means and topic modelling. IET Networks, Vol. 10, No. 6, pp. 278–294. DOI: 10.1049/ntw2.12017.

2. **Ali, T., Omar, B., Soulaimane, K. (2022).** Analyzing tourism reviews using an lda topic-based sentiment analysis approach. MethodsX, Vol. 9, pp. 101894. DOI: 10.1016/j.mex.2022.101894.

3. **Chaganti, S. Y., Nanda, I., Pandi, K. R., Prudhvith, T. G., Kumar, N. (2020).** Image classification using SVM and CNN. International Conference on Computer Science, Engineering and Applications, pp. 1–5. DOI: 10.1109/iccsea49143.2020.9132851.

4. **Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., Hassanien, A. E. (2020).** Sentiment analysis of COVID-19 tweets by deep learning classifiers — A study to show how popularity is affecting accuracy in social media. Applied Soft Computing, Vol. 97, pp. 106754. DOI: 10.1016/j.asoc.2020.106754.

5. **Eurostat (2022).** Consumption of online news rises in popularity. Eurostat News. ec.europa.eu/eurostat/en/web/products-eurostat-news/-/ddn-20220824-1.

6. **Hamborg, F., Donnay, K., Gipp. B. (2018).** Automated identification of media bias in news articles: An interdisciplinary literature review. International Journal on Digital Libraries, Vol. 20, No. 4, pp. 391–415. DOI: 10.1007/s00799-018-0261-y.

7. **Hammad, R., Barhoush, M., Abed-Alguni, B. H. (2020).** A semantic-based approach for managing healthcare big data: A survey. Journal of Healthcare Engineering, Vol. 2020, pp. 1–12. DOI: 10.1155/2020/8865808.

8. **Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y. (2017).** Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, Vol. 2, No. 4, pp. 230–243. DOI: 10.1136/svn-2017-000101.

9. **Khurana, D., Koli, A., Khatter, K., Singh, S. (2022).** Natural language processing: State of the art, current trends and challenges. Multimedia Tools and Applications, Vol. 82, No. 3, pp. 3713–3744. DOI: 10.1007/s11042-022-13428-4.

10. **Kirill, Y., Mihail, I. G., Sanzhar, M., Rustam, M., Olga, F., Ravil, M. (2020).** Propaganda identification using topic modelling. Procedia Computer Science, Vol. 178, pp. 205–212. DOI: 10.1016/j.procs.2020.11.022.

11. **Kulkarni, R. (2022).** A million news headlines. Harvard Dataverse. dataverse.harvard.edu/

12. **Kurenkov, A. (2020).** Lessons from the PULSE model and discussion. The Gradient thegradient.pub/pulse-lessons/.

13. **Liu, B. (2012).** Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, Springer International Publishing. DOI: 10.1007/978-3-031-02145-9.

14. **Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., Trajanov, D. (2020).** Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE Access, Vol. 8, pp. 131662–131682. DOI: 10.1109/access.2020.3009626.

15. **Mohammad, S. M., Turney, P. D. (2013).** NRC emotion lexicon. National Research Council Canada Publications Record. DOI: 10.4224/21270984.

16. **Muhammad, Z., Jailani, N. A. J., Leh, N. A. M., Hamid, S. A. (2022).** Classification of drinking water quality using support vector machine (SVM) algorithm. Proceedings of the IEEE 12th International Conference on Control System, Computing and Engineering, pp. 75–80. DOI: 10.1109/iccsce54767.2022.9935657.

17. **Nzotta, C. (2023).** A quick history of natural language processing. Aveni aveni.ai/history-of-natural-language-processing/.

18. **Pang, B., Lee, L. (2008).** Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, Vol. 2, No. 1–2, pp. 1–135. DOI: 10.1561/1500000011.

19. **Pang, B., Lee, L., Vaithyanathan, S. (2002).** Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vol. 10, pp. 79–86. DOI: 10.3115/1118693.1118704.

20. **Pew Research Center (2021).** Social media use in 2021. Pew Research Center. www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

21. **Ponn-Felciah, M. L., Anbuselvi, R. (2016).** Smartphone product review sentiment analysis using logistic regression. International Journal of Circuit Theory and Applications, Vol. 9, No. 26, pp. 343–349.

22. **Sharma, D. N., Shankar, D. P., Raj, M. R., Dalwadi, M. C. (2022).** Sentiment analysis for amazon product reviews using logistic regression model. Journal of Development Economics and Management Research Studies, Vol. 9, No. 11, pp. 29–42. DOI: 10.53422/jdms.2022.91104.

23. **Singh, N. K., Tomar, D. S., Sangaiah, A. K. (2018).** Sentiment analysis: A review and comparative analysis over social media. Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No. 1, pp. 97–117. DOI: 10.1007/s12652-018-0862-8.

24. **Valenti, A. P., Chita-Tegmark, M., Tickle-Degnen, L., Bock, A. W., Scheutz, M. J. (2019).** Using topic modeling to infer the emotional state of people living with parkinson's disease. Assistive Technology, Vol. 33, No. 3, pp. 136–145. DOI: 10.1080/10400435.2019.1623342.

Article received on 15/04/2024; accepted on 23/06/2024.
**Corresponding author is Fernando Perez-Tellez.*