

Análisis Sentimental en los Medios de Comunicación Colombianos para la Evaluación de Tendencias Políticas Nacionales

Julian D. Osorio , *Universidad Nacional de Colombia*

Resumen— Este proyecto propone llevar a cabo un análisis sentimental de los principales medios de comunicación colombianos con el fin de evaluar las tendencias políticas nacionales. Se centra en la problemática del posible sesgo y la manipulación de la opinión pública por parte de estos medios, cuya influencia y alcance pueden tener un impacto significativo en la formación de la opinión política. Mediante el análisis de sentimientos, se busca comprender mejor las tendencias impuestas por estos medios y proporcionar una base para la toma de decisiones políticas más informadas y pertinentes.

Abstract— This project proposes to carry out a sentimental analysis of the main Colombian media in order to evaluate national political trends. It focuses on the problem of possible bias and manipulation of public opinion by these media, whose influence and reach can have a significant impact on the formation of political opinion. Through sentiment analysis, we seek to better understand the trends imposed by these media and provide a basis for making more informed and relevant political decisions.

Index Terms— Análisis sentimental, medios de comunicación, tendencias políticas, opinión Pública, procesamiento de Lenguaje Natural.

I. INTRODUCCIÓN

LOS medios de comunicación desempeñan un papel fundamental en la sociedad contemporánea al influir en la opinión pública y dar forma a las tendencias en diversos ámbitos de interés general. En el contexto colombiano, medios como El Tiempo, El Espectador y Semana tienen una presencia destacada y un alcance significativo, lo que les otorga un poder considerable para moldear la percepción de la ciudadanía sobre temas políticos [1]. Así pues, los medios han empleado sus recursos para la promoción de agendas y candidatos políticos particulares [2]; por ello existe la preocupación sobre la posible parcialización y la manipulación de la información por parte de estos medios [3].

Para abordar esta problemática, es crucial contar con herramientas y enfoques analíticos que nos permitan comprender de manera objetiva y sistemática las tendencias políticas promovidas por estos medios. En este sentido, el análisis sentimental emerge como una metodología poderosa que permite evaluar el contenido de las noticias [4]. Es un conjunto de técnicas que utilizan el procesamiento de lenguaje natural; así pues, es posible aplicar diversos métodos del aprendizaje de máquina, como lo son la Red neuronal recurrente (RNN), Máquina de vectores de soporte (SVM), Bosque aleatorio (RF) y la Red Convolutiva Naive Bayes (CNN), esto mediante el uso de diversas librerías en el lenguaje de programación Python. Lo anterior con el fin de determinar la actitud, los sentimientos y las emociones de un texto. En el contexto de los medios

de comunicación, el análisis sentimental puede revelar cómo la forma en que se presenta una noticia puede influir en la percepción del público [5].

Basándose en investigaciones recientes en el campo del análisis sentimental de los medios de comunicación y las noticias, se espera profundizar en la comprensión de cómo los medios pueden influir en las tendencias políticas dentro del contexto nacional.

II. TRABAJOS RELACIONADOS

El análisis sentimental es un método pujante que constantemente goza de avances significativos conforme se desarrollan las técnicas del aprendizaje profundo. Diversos investigadores han contribuido, en particular, a la aplicación del análisis sentimental en las noticias utilizando diferentes enfoques. Se presenta ahora una breve discusión acerca del trabajo realizado con relación a esta aplicación puntual.

Taj, Shaikh y Meghji presentan una metodología basada en el análisis del léxico de la noticia. Donde se utilizó la técnica de frecuencia de términos de documentos inversa (TF-IDF) para identificar palabras importantes en los documentos y asignar pesos según la aparición de palabras; y con ello determinar la polaridad de esta misma. Con ello lograron sintetizar artículos del conjunto de datos de la BBC, que brindan información sobre los tonos emocionales presentes en función de diferentes categorías de contenido de noticias. Denotando exitosamente la diferencia de naturaleza que puede existir entre artículos deportivos o políticos, por mencionar solo un ejemplo [6].

Balahur et al. separan opiniones positivas o negativas de buenas o malas noticias; bajo la aplicación de diversos diccionarios de sentimientos [7]; esto siguiendo la conceptualización de la minería de opiniones como una disciplina encargada de la recuperación de información y la lingüística computacional donde no se aborda el tema del que trata un documento per se, sino de la opinión que expresa [8]. Similarmente Samuels y Mcgonigal distinguen entre opiniones con sentimiento y opiniones sin sentimiento y entre opiniones con sentimiento expresadas directa e indirectamente, donde es posible percibir la favorabilidad contra la desfavorabilidad, el apoyo contra la oposición, la crítica contra el aprecio o el agrado contra el disgusto; que no es explícito [9].

Bajo este orden de ideas, Islam, Ashraf, Abir y Mottalib propusieron un enfoque para clasificar noticias en línea. Se realizó un análisis de sentimientos a nivel de oración utilizando un diccionario dinámico para determinar la polaridad del sentimiento. El proceso de clasificación de artículos de noticias incluyó la selección de un artículo en línea, la extracción de oraciones, la búsqueda de palabras positivas en esas oraciones

y la combinación de las polaridades para obtener la polaridad final del artículo [10]. Para ello Python ofrece diversos recursos y librerías tanto para extraer como analizar los textos; Agarwal et al. aplican minería de opinión usando Natural Language Toolkit (NLTK) que brinda diversos algoritmos de procesamiento de lenguaje natural [11]. Para un enfoque basado en léxico, Meyer, Bikdash y Dai utilizaron el modelo Bag of Words (BOW) junto con General Inquirer Lexicon (H4N) para determinar las polaridades de los sentimientos [12].

Otro método para hallar la polarización de un artículo completo, es propuesta por Shirsat, Jagdale and Deshmukh donde se realiza la preparación del texto, que incluyó tokenización, eliminación de palabras vacías y reducción de palabras a su raíz, se realizó un proceso posterior en el artículo de noticias completo. En esta etapa, se asignó un puntaje de sentimiento al artículo, categorizando como positivo, negativo o neutral [13].

Con la introducción de las redes neuronales, Souma, Vondenska y Aoyama utilizan redes neuronales recurrentes (RNN) con unidades LSTM para entrenar el modelo en la predicción de sentimientos de noticias financieras [14]. Attila y Kiss utiliza una variedad de herramientas para el análisis de sentimiento, incluyendo BERT, VADER, TextBlob y RNN; e compararon los resultados de estas herramientas y se evaluó su precisión en la determinación de valores de polaridad [15].

III. METODOLOGÍA

Se aplican modelos de análisis BERT, VADER y Flair para su posterior evaluación y comparación de resultados. Para ello se sigue la secuencia de a continuación.

1. Preprocesamiento de Datos.

- Se recolecta un conjunto de datos de noticias relevantes para el análisis sentimental.
- Se realiza la limpieza de los textos, incluyendo la eliminación de signos de puntuación, palabras vacías y cualquier otro ruido o información irrelevante.
- Se traducen los textos al inglés para mayor facilidad de análisis.

2. Modelado.

- Se utiliza la implementación de TextBlob, BERT, VADER y Flair (con unidades LSTM) disponibles en Python, utilizando bibliotecas de HuggingFace.
- Cada modelo se entrena con los datos preprocesados y se ajustan los parámetros según sea necesario para optimizar el rendimiento.

3. Evaluación y Análisis Comparativo.

- Se estudian los resultados de cada modelo y se determina su eficiencia en la predicción del sentimiento de las noticias.
- Se identifica qué modelo proporciona las predicciones más precisas y coherentes, con respecto a las noticias estudiadas; con la mejor velocidad de procesamiento.

4. Selección y Validación.

- Basándose en los resultados de la evaluación y el análisis comparativo, se selecciona el modelo que mejor se

adapte a las necesidades y requisitos específicos para el análisis sentimental de noticias.

- Se valida el modelo seleccionado con un conjunto de datos extraídos de medios de comunicación puntuales en función de una temática relevante dentro de la coyuntura nacional.

IV. DISEÑO

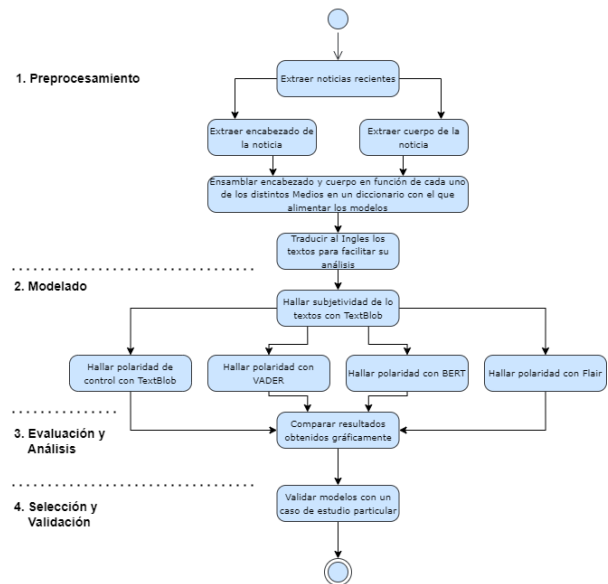


Figura 1. Flujo de Análisis Sentimental de Noticias.

V. DESARROLLO

V-A. Extracción y Procesamiento de Noticias

Primeramente se extraen las url de las noticias más recientes relativas a la actualidad de la política Nacional, tales que los encabezados contengan palabras clave definidas.

```

1 client = gnewsclient.NewsClient(
2     location='Colombia', language='spanish',
3     topic='Nation', use_opengraph=True,
4     max_results = 500
5 )
6 noticias = client.get_news()
7
8 p_clave = [] # aquí se definen las
9     palabras clave a considerar
10
11 noticias = [ noticia for noticia in
12     noticias if any( palabra.lower() in
13     noticia['title'].lower() for palabra
14     in p_clave ) ]
  
```

Las url de las noticias se almacenan en una lista y son procesadas con una función `get_cuerpo(noticias)`, que retorna un diccionario con los encabezados y texto de la noticia en función de cada medio de comunicación identificado.

V-B. Modelamiento

Una vez se tiene el diccionario *cuerpo*, con los contenidos de las noticias, este es traducido al Inglés, lenguaje en el cual se facilita el uso de los modelos, al ser más fácilmente reconocido.

```
1 traductor = Translator(from_lang='es',
2   ↪ to_lang='en')
3 for key in cuerpo.keys():
4     for sub_key in cuerpo[key].keys():
5
6         for i in range(
7             ↪ len(cuerpo[key][sub_key]) ):
8             texto = cuerpo[key][sub_key][i]
9             segmentos = [texto[i:i+500] for i in
10                ↪ range(0, len(texto), 500)]
11
12             traduccion = ''.join(
13                 ↪ traductor.translate(segmento)
14                 ↪ for segmento in segmentos )
15
16             cuerpo[key][sub_key][i] = traduccion
```

Una vez se han traducido los textos, estos son analizados con los diferentes modelos, los cuales son implementados en funciones independientes que retornan la polaridad promedio de encabezados y textos en función de cada medio de comunicación analizado.

- polaridad_blob(cuerpo)
- polaridad_vader(cuerpo)
- polaridad_bert(cuerpo)
- polaridad_flair(cuerpo)

A su vez se determina la subjetividad promedio de encabezados y textos con TextBlob implementando una función subjetividad(cuerpo).

V-C. Evaluación

Finalmente los resultados de Polaridad y Subjetividad se convierten en un DataFrame para facilitar su visualización. Donde se definen 3 diccionarios, *encbz-df* y *txt-df* que contiene la polaridad en función de cada medio de encabezados y textos respectivamente. A su vez *subj-df* que similarmente, contiene la subjetividad de encabezados y textos en función de cada Medio.

```
1 for key in cuerpo.keys():
2     encbz_df['Medio'].append(key)
3     txt_df['Medio'].append(key)
4
5     subj_df['Medio'].append(key)
6     subj_df['Encabezados'].append(
7         ↪ subjs[key]['encabezados'] )
8     subj_df['Textos'].append(
9         ↪ subjs[key]['textos'] )
10
11     for dic, model in zip(resultados,
12        ↪ models):
13         txt_df[model].append(
14             ↪ dic[key]['textos'] )
15         encbz_df[model].append(
16             ↪ dic[key]['encabezados'] )
```

```
12 encbz_df = pd.DataFrame(encbz_df)
13 txt_df = pd.DataFrame(txt_df)
14 subj_df = pd.DataFrame(subj_df)
```

De forma análoga para la evaluación de los tiempos de ejecución de los modelos.

```
1 time_df = {
2     'Blob - Subj' : [subjs_t],
3     'Blob - Polrd' : [blob_t],
4     'VADER' : [vader_t],
5     'BERT' : [bert_t],
6     'Flair' : [flair_t],
7 }
8
9 time_df = pd.DataFrame(time_df)
```

VI. IMPLEMENTACIÓN

El programa se implementa extrayendo noticias desde GoogleNews. Para este caso de estudio se hará un análisis sencillo de las posibles tendencias de los Medios que han cubierto recientemente, a fecha de escrito este trabajo, temas relativos al Gobierno Nacional actual. Cabe aclarar que las noticias extraídas y por ende los resultados obtenidos, variarán en función de la fecha en que dichas noticias sean extraídas, pues el programa se centra en estudiar la actualidad nacional, analizando las noticias más recientes del último mes de los principales medios de comunicación.

Para ello se definen palabras clave relacionadas a dichos temas para acotar la búsqueda de noticias.

```
1 '''
2 Posibles palabras clave
3
4 Política, Gobierno, Presidente, Ministro,
5   ↪ Senado, Congreso, Partido, Elecciones,
6   ↪ Reforma, Ley,
7   ↪ Constitución, Democracia, Corrupción,
8   ↪ Protesta, Manifestación, Voto,
9   ↪ Campaña, Candidato, Derechos,
10  Política exterior, Política interna,
11  Economía, Inflación, Desempleo,
12  Impuestos, Salud, Educación,
13  Seguridad
14 '''
15
16 p_clave = ['presidente', 'gobierno',
17            ↪ 'constituyente', 'petro',
18            ↪ 'mandatario', 'reforma', 'ley',
19            ↪ 'corrupción']
```

Con lo que se pueden obtener noticias como la siguiente.

Noticias de Cali, Valle y Colombia - Periódico: Diario El País
 Presidente Petro llega a Cali este 18 de mayo para "gran movilización por la vida"
 El presidente Gustavo Petro llegará a Cali el próximo viernes, 18 de mayo, para "una gran movilización por la vida y el manda ...

Figura 2. Noticia extraída de Diario El País.

Posteriormente el flujo de análisis sentimental es aplicado a las noticias; con lo que se determina subjetividad y polaridad de los medios estudiados en este caso. Los resultados obtenidos son organizados para su fácil visualización:

Medio	Encabezados	Textos
Semana	0.375	0.3291
El País	0.750	0.5042

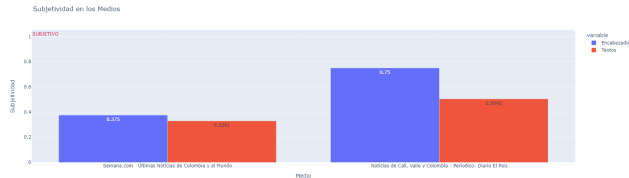


Figura 3. Subjetividad promedio para encabezados y textos.

Medio	Blob	VADER	BERT	Flair
Semana	0.15	0.0431	-0.9678	-0.7699
El País	0.80	0.6249	0.9974	0.7867

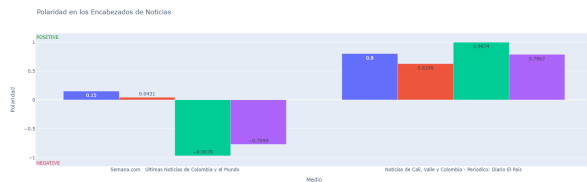


Figura 4. Polaridad promedio para encabezados.

Medio	Blob	VADER	BERT	Flair
Semana	-0.0087	-0.0076	-0.3663	-0.9474
El País	0.2133	0.9729	0.9656	0.8518



Figura 5. Polaridad promedio para textos.

Blob - Subj	Blob - Polrd	VADER	BERT	Flair
0.024	0.030	0.044	8.260	7.761

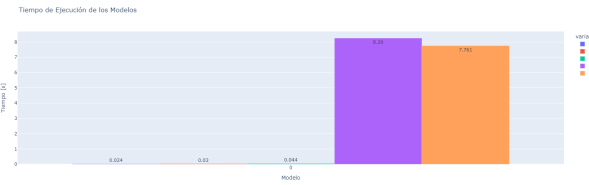


Figura 6. Tiempo de ejecución de los modelos en Segundos.

VII. RESULTADOS Y DISCUSIÓN

Se puede observar que para la *Revista Semana*, una editorial usualmente más favorable a la derecha del país, se obtiene una tendencia de polaridad inclinada al sentimiento negativo en las noticias acerca de un Gobierno fuertemente inclinado a las ideologías de izquierda, lo que mejor se aprecia en BERT y Flair, donde se exhibe una alta polaridad promedio en Encabezados, que es natural, al ser aquellos destinados a llamar la atención del lector y una polaridad menor en los textos en sí de las noticias. Aún así, se observa una subjetividad moderada tanto en encabezados como textos.

Por otro lado, para el *Diario el País*, de ideología más afín a la izquierda política, la tendencia es claramente positiva, con una subjetividad alta en encabezados y textos siempre mayor a 0,5. Como particularidad de este medio de comunicación, exceptuando TextBlob, el resto de modelos registra una polaridad similarmente elevada, tanto en encabezados como en textos, lo que es contrario a la intuición de tener un encabezado más polarizado y atractivo, acompañado de un texto más reservado y sutil.

En todo caso, se aprecia que los resultados producidos por BERT y Flair varían considerablemente a los vistos en TextBlob y VADER, en particular para aquellos de Revista Semana; exhibiendo una cercanía mayor a la realidad nacional, por lo que se puede inferir, que existe una mayor precisión en los modelos de redes neuronales, lo que coincide con los resultados vistos en [14].

Para este caso de estudio sencillo, se observa un rendimiento consistente y similar entre los modelos basados en diccionarios TextBlob y VADER, teniendo una velocidad significativa mayor a los modelos basados en redes neuronales BERT y Flair, tal y como se observó en [9]. A su vez los resultados obtenidos para modelos basados en diccionarios varían ligeramente entre ellos mismos, similarmente para modelos de redes neuronales. Así mismo se puede apreciar en [10], aunque en este caso particular, si bien la polaridad resultante para encabezados de TextBlob y VADER son bastante cercanas, la polaridad para textos denota mayor variación entre ambos modelos.

Considerando que los resultados obtenidos en BERT y Flair fueron bastante similares, así como el tiempo de ejecución requerido para estos dos modelos, ambos demuestran ser opciones viables para el análisis de noticias.

De lo anterior y los resultados vistos se puede resaltar que modelos basados en redes neuronales, como BERT y Flair,

pueden ser más pertinentes para el análisis de sentimientos en textos complejos, siendo más capaces de capturar el contexto y la semántica que los hace más adecuados para aplicaciones donde se necesita un análisis de sentimientos en textos largos o ambiguos, como noticias, artículos de opinión o comentarios extensos. Tales como aquellos estudiados en este caso particular o aquellos vistos en [13].

Aún así, los modelos basados en diccionarios, como TextBlob y VADER, pueden ser útiles para el análisis de sentimientos sencillo y eficiente. Estos modelos son fáciles de implementar y ofrecen resultados rápidos, lo que los hace ideales para aplicaciones donde se requiere un análisis de sentimientos rápido y sencillo, como la clasificación de comentarios en redes sociales o la evaluación de reseñas de productos. Adicionalmente son útiles para aplicaciones con recursos limitados; debido a su menor complejidad computacional, los modelos basados en diccionarios son más adecuados para aplicaciones en entornos con recursos limitados, como dispositivos móviles o sistemas embebidos. Sirviendo de referencia los textos estudiados en [6].

VIII. CONCLUSIONES

- Los modelos más rápidos en ejecutarse fueron TextBlob con 0,024 s y 0,030 s para análisis de subjetividad y polaridad respectivamente y VADER con 0,044 s; como era de esperarse al ser un modelos basado en diccionarios. A pesar de ello presentaron ambos un grado de precisión menor. Por lo que estos modelos resultan de mayor utilidad para el análisis de textos de menor complejidad cuando se necesite de mayor eficiencia y menor demanda computacional.
- Los modelos más precisos fueron BERT y Flair basados en redes neuronales. Teniendo BERT un puntaje de polaridad promedio para encabezados y textos de $-0,9678$ y $-0,3663$ para Semana y de $0,9974$ y $0,9656$ para El País. Flair registró un puntaje de polaridad de $-0,7699$ y $-0,9474$ para Semana y de $0,7867$ y $0,8518$ para El País. Lo que evidencia un uso elevado del discurso parcializado y sentimental en algunos de los principales medios de comunicación del país.
- Tanto BERT como Flair tomaron considerablemente más tiempo en ejecutarse, en comparación con los modelos basados en diccionarios. Tomando $8,260$ s y $7,761$ s respectivamente. Aún así, los tiempos de ejecución registrados no son inviablemente elevados y gracias a dicha precisión observada en estos modelos, tanto BERT como Flair resultan más pertinentes para textos complejos y de más sutilezas, como pueden serlo las noticias; siempre que se cuente con los recursos computacionales adecuados.
- Se puede apreciar que algunos de los medios nacionales apelan al uso de un discurso que va de lo moderado a lo altamente subjetivo. Obteniendo Semana una puntuación de subjetividad promedio para encabezados y textos de $0,375$ y $0,3291$. A su vez, El País obtuvo una puntuación de $0,750$ y $0,5042$.
- Se observa que en general la polaridad de encabezados es mayor que la de los textos. Esto puede ocurrir debido a que los encabezados se disponen para ser llamativos y captar la atención del lector para que este lea la noticia y le genere tráfico a los medios de comunicación.

REFERENCIAS

- [1] R. Pérez and L. Gómez, 'El Papel de la Prensa en la Formación de la Opinión Pública sobre Temas Políticos en Colombia', *Revista Latinoamericana de Comunicación*, vol. 8, no. 2, pp. 112-129, 2021.
- [2] M. Mathur, A. Agarwal and S. Gupta, "A Behavioural Analysis of US Election 2020 Using Deep Learning," 2022 IEEE 6th Conference on Information and Communication Technology (CICT), Gwalior, India, 2022, pp. 1-5, doi: 10.1109/CICT56698.2022.9997942.
- [3] C. Vargas and P. Herrera, "Percepciones y Actitudes de los Ciudadanos Colombianos hacia los Medios de Comunicación y su Influencia en la Opinión Pública," *Revista de Estudios de Comunicación*, vol. 18, no. 2, pp. 56-73, 2019.
- [4] K. Thiagarajan, A. Bansal and N. Muthurasu, "Sentimental Analysis on Media," 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET), B G NAGARA, India, 2023, pp. 1-5, doi: 10.1109/ICRASET59632.2023.10420291.
- [5] L. Y. N, R. N. N, S. K, A. R. K P and G. S, "Sentimental Analysis on Social Media Dataset Using Different Algorithms in Machine Learning," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAIS), Trichy, India, 2022, pp. 350-357, doi: 10.1109/ICAIS55157.2022.10011134.
- [6] S. Taj, B. B. Shaikh and A. Fatemah Meghji, "Sentiment Analysis of News Articles: A Lexicon based Approach," 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2019, pp. 1-5, doi: 10.1109/ICO-MET.2019.8673428.
- [7] A. Balahur et al., "Sentiment Analysis in the News," 2013. [Online]. Available: arXiv:1309.6202 [cs.CL].
- [8] Esuli, A. and F. Sebastiani. (2006). SentiWordNet: A Publicly Available Resource for Opinion Mining. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2006, Italy*. 2006.
- [9] Antony Samuels y John Mcgonical, "News Sentiment Analysis," arXiv:2007.02238 [cs.CL], 2020.
- [10] M. U. Islam, F. B. Ashraf, A. I. Abir and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, 2017, pp. 1-5. doi: 10.1109/ICCIT-TECHN.2017.8281777
- [11] A. Agarwal, V. Sharma, G. Sikka and R. Dhir, "Opinion mining of news headlines using SentiWordNet," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp.1-5. doi: 10.1109/CDAN.2016.7570949
- [12] B. Meyer, M. Biddash, and X. Dai, "Fine-grained financial news sentiment analysis," *SoutheastCon 2017*, 2017.
- [13] V. S. Shirsat, R. S. Jagdale and S. N. Deshmukh, "Document Level Sentiment Analysis from News Articles," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-4.
- [14] W. Souma, I. Vodenska and H. Aoyama, "Enhanced News Sentiment Analysis Using Deep Learning Methods," *Journal of Computational Social Science*, vol. 2, pp. 33-46, 2019.
- [15] Nemes, Kiss, A. (2021). "Prediction of stock values changes using sentiment analysis of stock news headlines." *Journal of Information and Telecommunication*, 5(3), 375-394.