CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17823

Unveiling New Insights From Textual Unstructured Big Data in Politics Through Deep Learning

Ufuk Caliskan[1], Angela Pappagallo[2], Francesco Ortame[2], Mauro Bruno[2], Francesco Pugliese[2]

[1] Deutsche Post & DHL, Germany, [2]Italian National Institute of Statistics, Italy.

How to cite: Caliskan, U.; Pappagallo, A.; Ortame, F; Bruno, M.; Pugliese, F. 2024. Unveiling New Insights From Textual Unstructured Big Data in Politics Through Deep Learnin. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17823

Abstract

Over the past decade, social media platforms have undergone significant and rapid expansion. One of the key challenges has been effectively analysing the vast amount of unstructured user-generated data they produce. This research delves into the analysis of Italian Twitter data through the application of advanced deep learning models across three primary objectives: text classification, sentiment analysis, and hate analysis. Five cutting-edge models are evaluated, each utilizing distinct word embeddings. Furthermore, this study investigates the effects of processing emojis and emoticons in Italian tweets on sentiment and hate analysis. We compare model performances and suggest optimized approaches for each task. Finally, we apply these methodologies to real-world Twitter data and present our findings through multiple graphs and statistical analyses. This study demonstrates the possibility of extracting new insights and novel information from unstructured textual Big Data in Politics.

Keywords: *politics, deep learning, artificial intelligence, big data, statistics, sentiment*

1. Introduction

Social media platforms, exemplified by Twitter's massive user base of 330 million monthly users generating 500 million daily tweets, offer rich 'Big Data' replete with opinions and sentiments (Leonowicz-Bukała et al., 2021). Recent advances empower researchers to extract insights, notably in politics, aiding in real-time sentiment analysis. This study delves into Italian tweets, employing deep learning techniques like Word2Vec and Fasttext embeddings, along with models like CNN, LSTM, and RCNN, for political classification, sentiment, and hate analysis. Model performance metrics like accuracy and F1-score are scrutinized, with the best approach applied to test Twitter data, showcased through graphical representations. This

Editorial Universitat Politècnica de València

research highlights deep learning's prowess in distilling insights from vast social media datasets (Catanese et al., 2023).

2. Methods

Natural language, essential for human communication, evolves over time without explicit rules like programming languages. Natural Language Processing (NLP), a subset of AI, focuses on enabling computers to understand, manipulate, and interpret human language (Bird et al., 2009). Artificial Neural Networks (ANNs) model biological nervous systems and comprise interconnected neurons across input, hidden, and output layers (Yegnanarayana, 2009). Convolutional Neural Networks (CNNs), initially for computer vision, excel in text classification by extracting features regardless of text position, using convolution and pooling layers (Li et al., 2021). Recurrent Neural Networks (RNNs) maintain state across steps, while Long Short-Term Memory (LSTM) networks address long-term dependency challenges by retaining information through specialized gate mechanisms (Hochreiter and Schmidhuber, 1997). Bidirectional LSTMs (BiLSTM) improve performance by processing input sequences in both directions, accessing forward and backward information at each step (Cui et al., 2018). Attention BiLSTM enhances focus on crucial input elements, aiding in sequence processing (Luo et al., 2018). Recurrent Convolutional Neural Networks (RCNNs) combine advantages of RNNs and CNNs, addressing bias towards later words and window size challenges (Siwei Lai et al., 2015). Word Embeddings represent words as low-dimensional continuous vectors, capturing semantic relationships (Xing et al., 2014). Word2vec, by Google, predicts words from context or context from words, with Skip-gram favored for its robust learning (Mikolov et al., 2013). FastText, an extension of skip-gram, overcomes word2vec limitations by employing character-level embeddings (Bojanowski et al., 2017).

3. Results

This section presents three distinct tasks: Politician Classification, Sentiment Analysis, and Hate Analysis. The objective is to compare the three methods described in the previous sections and construct the optimal model for each task.

3.1 Politics Classifier

The main objective of this task is to create a brief text classifier using official tweets from five prominent Italian politicians. The tweets have been obtained through the Twitter API. To develop the classifier, numerous DL methods have been employed and compared. The best model has been then used to classify tweets from other users, including journalists and newspapers. The main idea of this work is that the language styles and keywords used by politicians can distinguish their texts and display similarities between them and their supporters.

The dataset used in this study includes tweets from the official accounts of five major Italian politicians: "*Giuseppe Conte*", *"Luigi Di Maio"*, *"Matteo Renzi"*, *"Matteo Salvini"*, and *"Nicola Zingaretti"*. A total of 13,541 tweets have been downloaded. URLs have been removed as they do not provide useful information for this task. Additionally, stopwords and non-alphanumeric characters have been deleted from the Training Set. Single characters and multiple white spaces have been removed, and every character has been converted to lowercase. After the preprocessing steps, the total number of tweets in the candidate Training Set is 8,376. As illustrated in Figure 1, the dataset is heavily imbalanced.

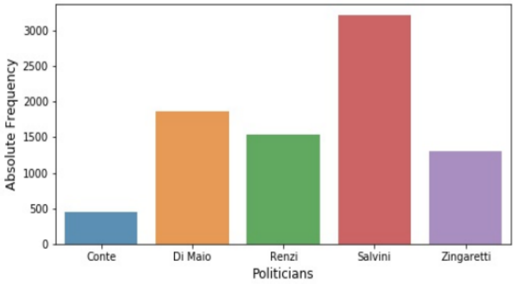
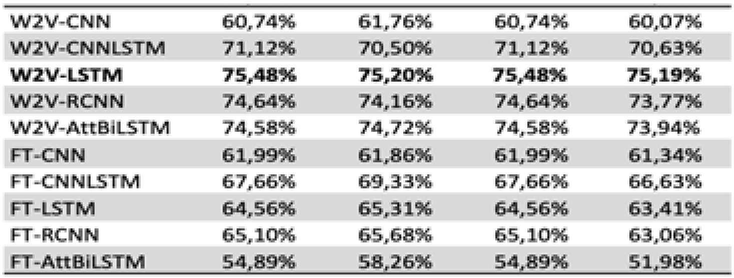


Figure 1. Dataset with Tweets of five major Italian Politicians.

"Giuseppe Conte" has the least Twitter activity with 456 tweets, while *"Matteo Salvini"* has the most with 3214 tweets. Table 1 shows the final performance metrics of the DL models considered by us, including accuracy, precision, recall, and F1-score. Due to the imbalanced condition of the training set, the F1-score is given higher importance.

Table 1. Accuracy, precision, recall and f1-score of all the Deep Learning methods.



W2V-CNN	60,74%	61,76%	60,74%	60,07%
W2V-CNNLSTM	71,12%	70,50%	71,12%	70,63%
W2V-LSTM	**75,48%**	**75,20%**	**75,48%**	**75,19%**
W2V-RCNN	74,64%	74,16%	74,64%	73,77%
W2V-AttBiLSTM	74,58%	74,72%	74,58%	73,94%
FT-CNN	61,99%	61,86%	61,99%	61,34%
FT-CNNLSTM	67,66%	69,33%	67,66%	66,63%
FT-LSTM	64,56%	65,31%	64,56%	63,41%
FT-RCNN	65,10%	65,68%	65,10%	63,06%
FT-AttBiLSTM	54,89%	58,26%	54,89%	51,98%

The table above presents a comparison of data preprocessing (embeddings) and Deep Neural Networks applied to the same Training Set. It is observed that W2V (Word2Vec) outperforms FastText. The CNN (1 Dimensional Convolutional Neural Network) model chosen for this task appears to be unsuitable. The combination CNN-LSTM model performs averagely. However, the LSTM model shows the best performance overall. The W2V-LSTM model has been chosen as the politician classifier, following AttBiLSTM (Bidirectional LSTM with Attention

Mechanisms) and RCNN (Recurrent Convolutional Neural Networks). Then, Official tweets related to Italian newspapers and journalists have been downloaded and classified to measure their closeness to the five politicians previously modelled during training. We have scraped 20,693 official tweets from seven Italian newspapers and 31,538 tweets from eleven Italian journalists. Both datasets have undergone the same preprocessing as the training stage. The W2V-LSTM model has been applied to both datasets in the inference stage. Contingency tables have been calculated and results are projected into a 2-dimensional space using Correspondence Analysis (CA). Canonical analysis (CA) is a multivariate statistical method used to visually represent dependencies in contingency tables. Figure 2-left shows the outcomes of the newspapers, while figure 2-right shows those of the journalists.

3.2. Sentiment Analysis

The objective of the Sentiment Analysis is to create a sentiment classifier for tweets and cross-reference the results with the politician classifier breakthroughs described in the previous section. The chosen Training Set is a combination of three different datasets: Sentipolc data [14], Happy Parents data [47], and Absita data [15]. The Sentipolc data comprises 9,410 Italian tweets, divided into 7,410 tweets for training and 2,000 tweets for the test set.

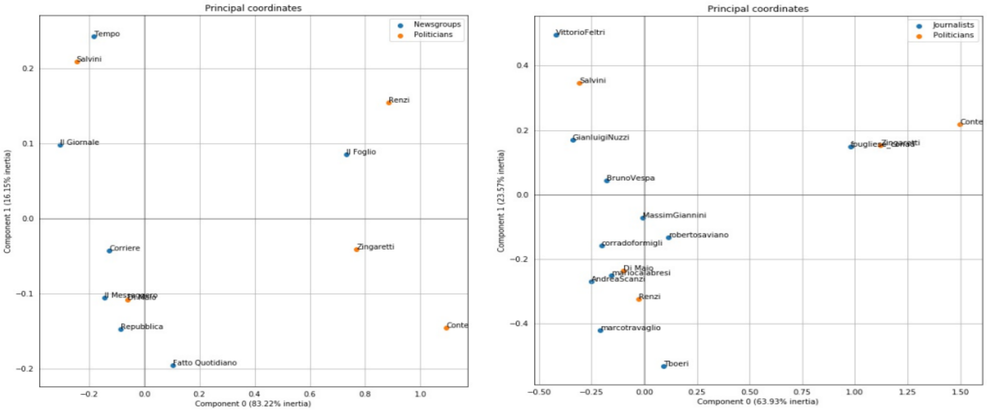


Figure 2. left) Visualization of Correspondence Analysis for Newspapers. The overlapping labels of the points in the bottom left are Il Messaggero for the blue point and Di Maio for the orange point; right) Visualization of Correspondence Analysis for Journalists. The overlapping labels of the points in the top right are fpugliese_conad for the blue point and Zingaretti for the orange point.

The dataset comprises both political and generic tweets, while the test data consists of tweets extracted using hashtags and keywords related to the socio-political topic #labuonascuola. Emoticons and emojis are included in the analysis by replacing them with the corresponding English textual representation using specific libraries (demoji [6]). After training the same deep neural networks as in the previous section, the W2V-RCNN method has been selected as the best model for this task. We have tested W2V-RCNN on tweets downloaded from politicians to classify their sentiment. Figure 3-left displays the distribution of the sentiment classes for each politician. Subsequently, the sentiment classifier and politics classifier have been jointly applied to tweets containing only the hashtags #primagliitaliani and #lilianasegre. The hashtag #primagliitaliani, meaning 'Italians first', has been actively promoted by Matteo Salvini. Instead, Liliana Segre is an Italian Holocaust survivor and a senator for life in Italy. A discussion about Liliana Segre's protection has been ongoing for years, and the hashtag #lilianasegre has become popular during autumn and winter 2019. We have downloaded 1,086 tweets with the hashtag #primagliitaliani between 18/11/2019 and 27/11/2019, and 1,994 tweets with the hashtag #lilianasegre between 21/11/2019 and 27/11/2019. The final classification of this sub-task consists of two steps. The text describes the distribution of political trends and sentiments in all these tweets. The results are presented in Figure 3-right, which shows that tweets with the hashtag #primagliitaliani are mostly associated with Matteo Salvini, followed by Luigi Di Maio. Classifications of Matteo Renzi and Nicola Zingaretti have been infrequent. Tweets classified as Matteo Salvini are mostly negative or neutral, while tweets classified as Luigi Di Maio are mostly neutral and positive. The tweets that include the hashtag #lilianasegre display a similar pattern, but with less impact from Matteo Salvini on the results.

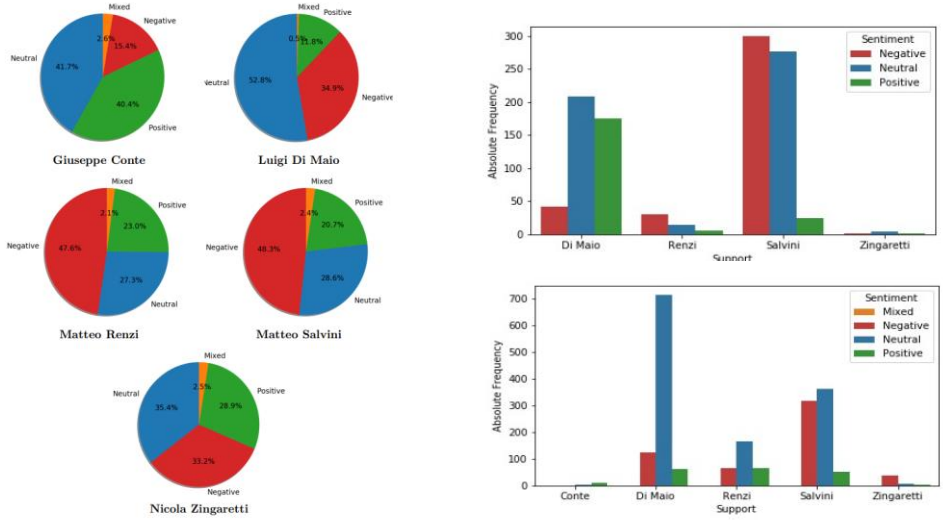


Figure 3.left) Sentiment of the tweets per politician; right) #primagliitaliani (up) and #lilianasegre (down) sentiment and politics simultaneous classification

3.3. Hate Analysis

The final task is building an Italian hate classifier, addressing the surge of online hate expression, notably on social media. Limited labeled datasets exist for hate classification, mirroring sentiment analysis challenges. Using Facebook comments and Twitter tweets, the study aims to analyze and apply findings to political figures. The hate recognition data utilized in this study originates from the EVALITA 2018 Hate Speech Detection Task, encompassing Facebook and Twitter datasets. For this datase, Facebook comments have been collected from public pages of Italian newspapers, politicians, artists, and groups, suspected to contain hateful content. A total of 99 posts have generated 17,576 comments, annotated by five students into classes: no hate, weak hate, and strong hate. Annotations have been were simplified for the EVALITA 2018 task, resulting in two classes: 0 for no hate and 1 for hate, with 4,000 posts in total. As with sentiment analysis, the preprocessing applied to the hate data has the steps, the emoticons translation included. After the training, the best model has been the W2V-RCNN, as in sentiment analysis. The hate classifier has been applied to tweets directed at the five politicians. For each politician, 2000 tweets with a minimum number words equal to three have been downloaded, in which they have been mentioned. The data have been collected between 25/11/2019 and 27/11/2019. After preprocessing, 5,937 tweets have been removed from the corpus of 10,000 tweets. By exploiting the trained hate classifier, each tweet from the remaining corpus of 4,063 newly download tweets has been classified as hateful or not hateful. In Figure 4 we depict the hate ratio of each politician during the reference period, which is calculated by dividing the number of hateful tweets by the total number of tweets.

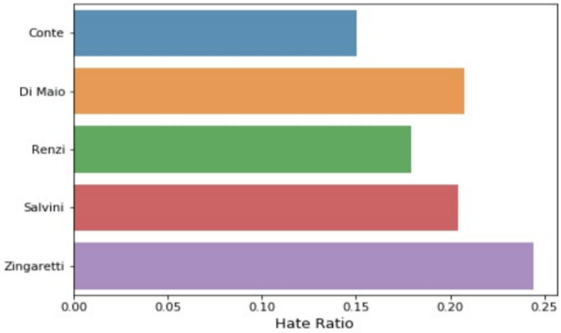


Figure 4 Distribution of the tweets with mentions

The hate classifier has identified that Nicola Zingaretti received the highest percentage of hateful tweets at 25%, followed by Matteo Salvini and Luigi Di Maio at 20%. Meanwhile, Matteo Renzi received 18% and Giuseppe Conte received 15% of hateful tweets in which they were mentioned. Finally, a graph was plotted to display the relationships between various politicians based on the negative sentiment expressed in tweets mentioning one politician by another, as shown in Figure 5. The thickness of the arrow indicates the strength of the negative relationship between two politicians or their supporters. This suggests that the thickness of the arrow may indicate a deterioration in the relationship between two politicians, while a slimmer arrow may indicate an improvement in their relationship.

4. Conclusions

This research conducts a comparative analysis of various methodologies to determine the most effective Deep Learning (DL) approach for classifying political tweets, conducting sentiment analysis, and detecting hate speech within tweets. Each task uses a different DL method specifically designed for tweet analysis. The W2V-LSTM method has been shown to perform well in classifying political tweets, achieving an F1-score of 75.19% on a dataset of tweets from five prominent Italian politicians. The study indicates that tweets from various users, especially shorter ones, demonstrate a writing style that is more closely associated with a particular political class, which affects classifier decisions. To improve classifier evaluation in future studies, it is recommended to establish a ground truth using tweets from users who express political opinions. Additionally, expanding the dataset to include more tweets and politicians is advised. The W2V-RCNN method outperforms others in sentiment analysis, achieving an F1-score of 77.58%. However, the inclusion of various datasets may skew findings, highlighting the need for a larger, labeled corpus for accurate evaluation. Additionally, further research is warranted to enhance analysis precision, particularly in processing emojis and emoticons and their wider range and equivalent translations. This work demonstrates the use of advanced Artificial Intelligence tools, specifically Deep Neural Networks (Deep Learning), to extract valuable insights from unstructured textual data such as tweets and short texts. These models provide significant benefits in modern data analysis as they can assist politicians (or individuals in general) in extracting additional information from textual data that would otherwise be unattainable from other sources. This creates innovative opportunities for data analysis in both Official Statistics and the analysis of the orientations of a reference population. In the future, more recent analysis models such as Transformers (Vanilla Transformer, Bert, GPT) could be

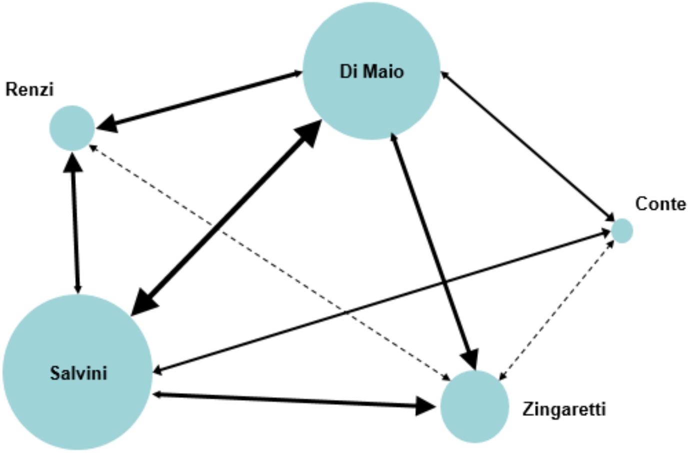


Figure 5 Distribution of the tweets with mentions

used. These models may achieve superior prediction metrics and provide better analysis of relationships between politicians with more sophisticated hate and sentiment analysis.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

Catanese, E., Bruno, M., Stefanelli, L., & Pugliese, F. (2023). Measuring Social Mood on Economy during Covid times: effects of retraining Supervised Deep Neural Networks.Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2018). Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. arXiv preprint arXiv:1801.02143.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Leonowicz-Bukała, I., Adamski, A., & Jupowicz-Ginalska, A. (2021). Twitter in Marketing Practice of the Religious Media. An Empirical Study on Catholic Weeklies in Poland. Religions, 12(6), 421.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 34(8), 1381-1388.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., & Qu, H. (2017, October). Understanding hidden memories of recurrent neural networks. In 2017 IEEE conference on visual analytics science and technology (VAST) (pp. 13-24). IEEE.

Raffel, C., & Ellis, D. P. (2015). Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint arXiv:1512.08756.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence, 2015.

Xing, C., Wang, D., Zhang, X., & Liu, C. (2014, December). Document classification with distributions of word vectors. In Signal and information processing association annual summit and conference (APSIPA), 2014 asia-pacific (pp. 1-5). IEEE.

Yegnanarayana, B. (2009). Artificial neural networks. PHI Learning Pvt. Ltd..