International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Sentiment Analysis of Italian and English Corpora of Internet News: A Comparison with Some Economic Trends

Luca Pavan
Institute of Foreign Languages, Vilnius University, Vilnius, Lithuania; Language Studies Center, Faculty of Creative Industries, Vilnius Tech, Vilnius, Lithuania; Department of Foreign Languages, Literary and Translation Studies, Vytautas Magnus University, Kaunas, Lithuania.
Corresponding Author: Luca Pavan, **E-mail**: pavan@panservice.it

| ABSTRACT
In this article, the sentiment analysis of several large Internet corpora made of Italian and English news is performed using a software written by the author, showing a possible connection with some economic trends. In this research, the news includes different topics (not necessarily financial news), and they are extrapolated from a large number of Internet newspapers. The software, already used in a previous article by the same author, is lexicon-based and makes use of scale points ranging from 0 to 100 to calculate an index of positivity in a text. The variation of sentiment tendency in the news corpora, calculated for a time period of several years, is later compared with some graphs showing some parameters of some economic trends, including the gross domestic product (GDP). It is found that the sentiment tendency of the news seems to have a relationship with the tendency of some economic trends that span the same time period. Positive growth of the economy per year seems connected with a positive variation in the index of positivity. Inversely, for a negative trend in the economy, the variation in the index of positivity is also negative. The article shows that, for various news topics, sentiment analysis can be useful to better understand some economic trends. For financial news, many studies show the possibility of predicting GDP growth through sentiment analysis. In this article, it is hypothesized that a prediction based on large news corpora including various topics could also be possible.

| KEYWORDS
Computational linguistics, Sentiment analysis, Internet media.

| ARTICLE DOI: 10.32996/ijllt.2022.5.5.17

1. Introduction

In recent years, many studies in the field of sentiment analysis have focused on the analysis of news. Today, large corpora of news are available on the Internet, giving the opportunity to analyze the opinions using a large amount of data. Articles from newspapers on the Internet are often freely available as archives, such as The BBC News dataset (Samuels, 2020). These archives can be analyzed for the polarity of sentiment (positive, negative, or neutral) or, more recently, the emotions behind the texts.

One of the tendencies in the studies of sentiment analysis is the prediction of economic trends through the examination of economic news. In recent years researchers have found a connection between financial news and the growth of economies, for example, by comparing the results coming from sentiment analysis with the gross domestic product (GDP) (Ashwin et al., 2021, pp. 4-7). To perform such an analysis, often lexicon-based software with a dictionary of words belonging to the field of business is used together with statistical methods (Ito et al., 2017, p. 896). Internet news is often filtered because much of them are considered irrelevant (Seki et al., 2022, p. 4). Other studies classify sentiment in the news more deeply, according to the author, the reader, and the text (Balahur, 2010, p. 2220). In other authors, the news is selected according to the presence or not of a token belonging to the field of the economy (Barbaglia et al., 2022, p. 3). In this article, the approach is different: sentiment analysis was applied to some news corpora in Italian and English language, but the news is extrapolated from Internet newspapers without taking into account their topics. The news corpora, spanning several years, were obtained from the Leipzig Corpora Collection

(https://corpora.uni-leipzig.de/en?corpusId=ita_news_2005-2009). Each file represents an entire year or a time period of several years. In this research, each text file contains 300,000 sentences. These corpora are made up of a random selection of sentences from a number of Internet newspapers. In this article, the texts include about 24 million words, while for the English news, the texts include roughly 60 million words. The goal is to show that for a large news corpus, it is possible to compare the sentiment coming from texts having various topics with some economic trends. This eventually also makes it possible to make predictions about the near-term state of the economy. As stated by Liu, acquiring public opinions has long been a huge business (Liu, 2012, p. 2). Behind news articles on the Internet, some opinions are also expressed and can be studied with sentiment software tools.

2. Method

To perform sentiment analysis, a lexicon-based software written by the author, called Psychoword, was used. The software was already used to analyze some literary works and was described in a previous article (Pavan, 2022). Psychoword is written in the C language and includes two versions: the first version is designed for the Italian language; the second version works for the English language. The Italian dictionary included in the software, created manually, includes approximately 30,000 sentiment words, partly derived from Porcu's dictionary (Porcu, 2016) and partly written by the author. These words are labeled using a scale of points from 0 to 100 (0 is fully negative, 50 is neutral, 100 is fully positive). In the English version, the dictionary is derived from that of Hu and Liu (https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html). The software also includes several valence shifters (Taboada et al., 2011, p. 269) and produces, among other things, an *index of positivity* (IP, which is given by the arithmetic mean of the sentiment scores). The dictionary does not include specific words belonging to economic fields. The aim of this research is, in fact, to analyze news where many topics are covered, so the dictionary does not need to be specifically designed for business.

The news, as they appear on their website, is digitally collected in corpora according to the language and presented in a number of files. Each file represents one or more years. In this article, several years of news are analyzed, from 2005 to 2020. For each year, using Psychoword, an index of positivity is calculated. Later, a graph is built over the years, showing the index of the positivity trend. Finally, this graph is compared to other graphs showing some economic trends. The aim is to observe whether the index of positivity tends to show similarities with some economic trends. Psychoword is also able to sort sentiment words according to their frequency, so the most frequent sentiment words in the news corpora are listed here.

3. Results and Discussion

Analyzing the Italian news according to time periods produced the results listed in Fig. 1. The indexes of positivity show, in general, a positive trend (above 50). The graph in Fig. 2 shows that IP decreased in 2010, following the economic crisis that started in 2009. In addition, in 2020, the IP decreases following the pandemic crisis.

IP – Index of positivity

PW – Number of positive words

NW – Number of negative words

Year	IP	PW	NW
2005-2009	55.83	134689	107026
2010	55.31	133289	108807
2019	58.82	144809	101586
2020	56.12	138966	109962

Fig. 1 – Results of the sentiment analysis of Italian news for several years

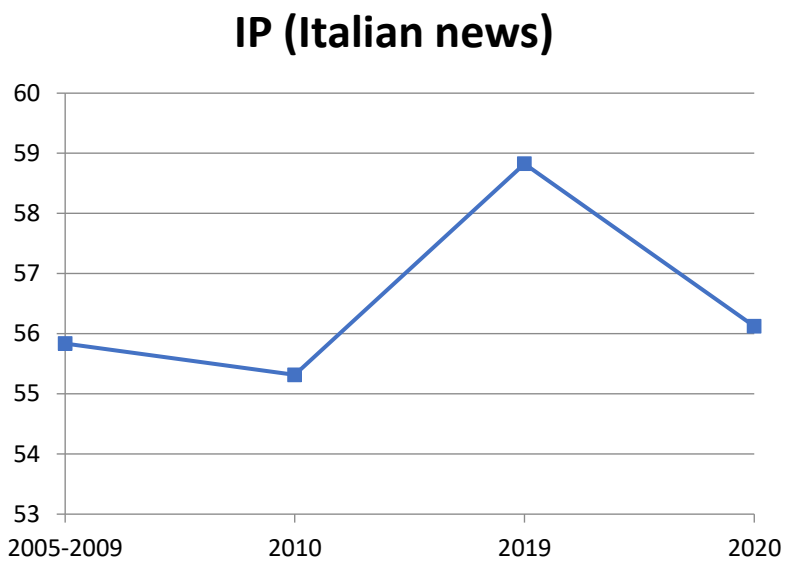

Fig. 2 – Index of positivity trend of Italian news for several years

It is possible to compare these results with a graph showing the Italian gross domestic product (PIL) in Fig. 3. Looking at the time period 2005-2020, it seems that the overall results of sentiment analysis are quite similar. However, the graph in Fig. 3 was done before the Ukrainian war, so the prediction appears too optimistic.

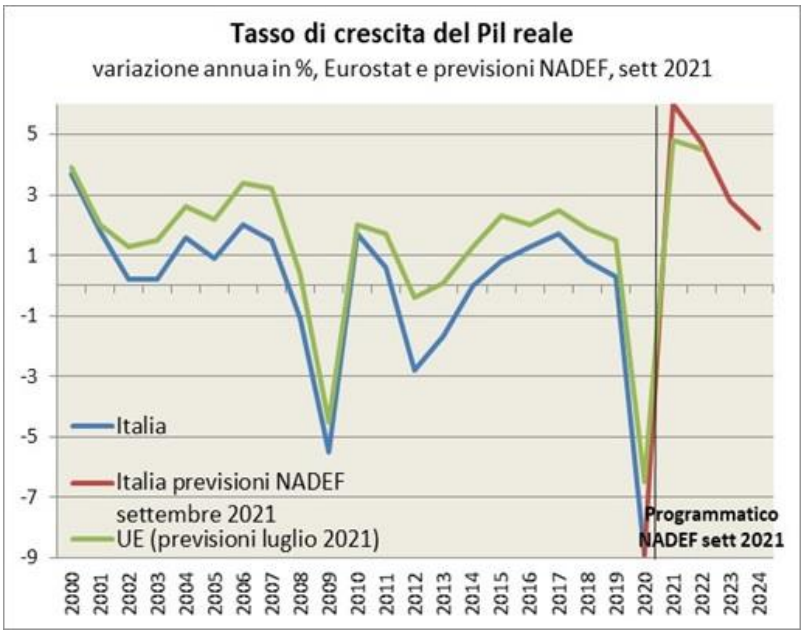

Fig. 3 – Italian GDP variation with prediction until 2024 (source: https://www.programmazioneeconomica.gov.it)

Psychoword is also able to list the most used sentiment words in the news. In the years 2005-2009, the most used positive word in the Italian news was *bene* (3288 times), and the most used negative word was *problema* (1976 times). Carrying out the same analysis for the other years gives similar results: *bene* and *problema* are the most used sentiment words also in the other years. The reason could be due to the language of newspapers, which is quite standardized, even if the news comes from different topics.

Comparing the IP tendency with the Italian import-export (Fig. 4) also reveals a similar trend.

Fig. 4 – Italian import-export (source: https://www.programmazioneeconomica.gov.it)

The case of English news is slightly different in comparison with the Italian news: in this case, the news is extrapolated from newspapers belonging to many countries having in common the English language. It is assumed that all these countries have similar economic trends because of their strong connections. These countries also represent a large part of the world economy, so in this article, there is a comparison between IP and the global world economic tendency. In Fig. 5, IPs for several years are listed.

Year	IP
2013	51,04
2014	50,16
2015	49,64
2016	58,27
2017	54,83
2018	55,36
2019	55,70
2020	54,25

Fig. 5 – Results of the sentiment analysis of English news for several years

English news shows a positive trend in IPs, except for the year 2015. In Fig. 6, there is a graph with the overall tendency.

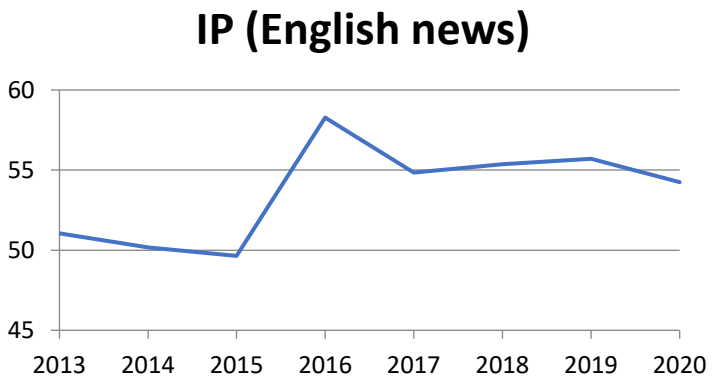

Fig. 6 – Index of positivity trend of English news for several years

Fig. 7 shows the trend of IP only for the years 2009-2010.

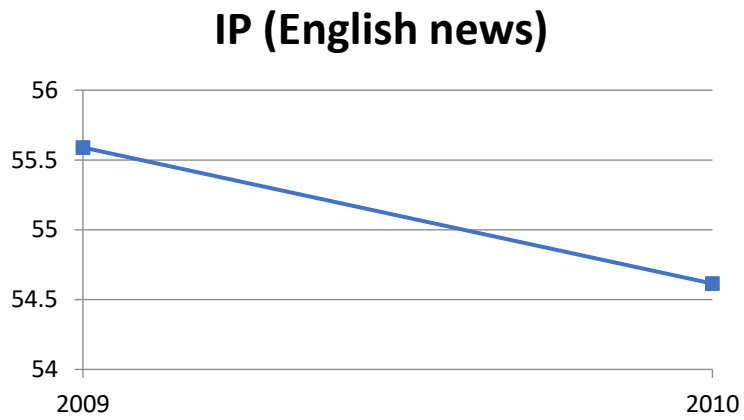

Fig. 7 – Index of positivity trend of English news for the years 2009-2010

Now it is possible to compare Fig. 6 and Fig. 7 with Fig. 8, which shows the world GDP for the years 1961-2020.

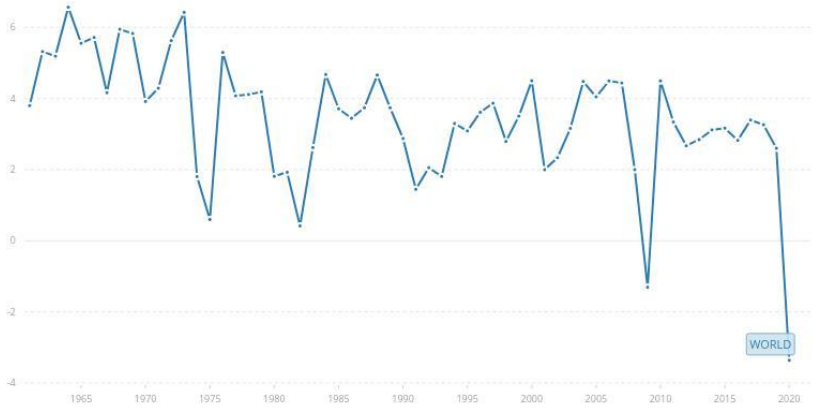

Fig. 8 – Gross domestic product (GDP) for the years 1961-2020 (source: https://data.worldbank.org)

When comparing the graphs, it is possible to see a match with the crisis of 2009, the pandemic crisis of 2020, and the increase in IP (and GDP) between 2017 and 2019. The most used positive sentiment words found with Psychoword are *good*, *well*, and *right*. The most used negative sentiment words are *hard*, *killed*, *death*, *issue*, *debt,* and *virus*.

Making predictions could also be possible to some extent, except in the case of a sudden crisis or, more recently, the Ukrainian war.

4. Conclusions

In recent years, sentiment analysis studies have focused extensively on social media and financial news. In this article, research was carried out on sentiment analysis of internet news belonging to various topics, showing the possibility of comparing the results with economic trends. It was found that there are similarities with the global economic trend for large corpora of news with different topics. To perform the analysis, a lexicon-based software written by the author in two languages (Italian and English) was used, including dictionaries not based on business sentiment words. This study shows that the general trend of opinions behind Internet news reflects, to some extent, economic trends. Many studies have already demonstrated a connection with the sentiment of financial news and the economy. The results of this research could be useful for future studies that aim to take into account global sentiment analysis of Internet news.

Funding: This research received no external funding.
Conflicts of Interest: The authors declares no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ashwin, J., Kalamara, E., and Saiz, L. (2021). Nowcasting Euro Area GDP with News Sentiment: A Tale of Two Crises *ECB Working Paper Series*. 2616.
[2] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J. (2010). Sentiment Analysis in the News. *Proceedings of the 7th International Conference on Language Resources and Evaluation*. 2216-2220. Valletta, Malta, 19-21 May.
[3] Barbaglia, L., Consoli, S., and Manzan, S. (2022). Forecasting with Economic News. *Journal of Business & Economic Statistics*, preprint.
[4] Ito, R., Izumi, K., Sakaji, H., and Suda, S. (2017). Lexicon Creation for Financial Sentiment Analysis Using Network Embedding. *Journal of Mathematical Finance*, 7. 896-907.
[5] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. San Rafael, USA: Morgan & Claypool Publishers.
[6] Pavan, L. (2022). A Survey of Some Italian Literature Works using Sentiment Analysis. *International Journal of Linguistics, Literature and Translation*, *5*(1). 117-121.
[7] Porcu, V. (2016). *Text mining e Sentiment Analysis con R.* Roma: Streetlib.
[8] Samuels, A., and Mcgonical, J. (2020). News Sentiment Analysis. Preprint, available on the Internet.
[9] Seki, K., Ikuta, Y., and Matsubayashi, Y. (2022). News-Based Business Sentiment and Its Properties as an Economic Index. *Information Processing and Management*, *59*(2). 1-16.
[10] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2):267–307.