Benchmarking Machine Learning for Sentiment Analysis: A Case Study of News Articles in Multiple Languages

Roberto Zanoli*
Fondazione Bruno Kessler

Alberto Lavelli**
Fondazione Bruno Kessler

Lorenza Romano†
Infojuice Srl

Verena Malfertheiner‡
Infojuice Srl

Pierluigi Casale§
OPIT - Open Institute of Technology

Sentiment analysis is the field of study that analyzes people's opinions and sentiments towards entities such as products, services and organizations. Brand reputation analysis, competitive intelligence and social network analysis are just a few areas that can benefit from sentiment analysis. Most studies on sentiment analysis have only focused on domains like product reviews and social network content, leaving sentiment inference in the news domain under-investigated. In this work, we use a case study of a company specialized in the analysis of brand reputation to evaluate machine learning models for sentiment analysis on multilingual news articles. Several models were tested, including traditional machine learning models like KNN, and transformer-based models like BERT, Llama and GPT. The implemented models were evaluated on a dataset of Italian, German and Ladin news articles annotated with their sentiment polarity. Overall, our experiments show state-of-the-art results and confirm the outcomes of previous studies, i.e. that sentiment analysis of news articles remains a complex task. Machine learning systems can support manual annotators in accelerating the annotation process. Our findings can provide a benchmark for researchers in natural language processing when performing sentiment analysis of news articles.

1. Introduction

The rapid growth of the Internet has enabled people to share their opinions on online review sites, blogs, and social networks. This feedback is invaluable to organizations seeking to understand public sentiment regarding their products and services.

* Fondazione Bruno Kessler - Via Sommarive, 18 - Povo 38123 Trento, Italy. E-mail: zanoli@fbk.eu
** Fondazione Bruno Kessler - Via Sommarive, 18 - Povo 38123 Trento, Italy. E-mail: lavelli@fbk.eu
† Infojuice Srl - Zona Produttiva Cardano,21 - Cornedo all'Isarco 39053, Italy
E-mail: lorenza.romano@infojuice.eu
‡ Infojuice Srl - Zona Produttiva Cardano,21 - Cornedo all'Isarco 39053, Italy
E-mail: verena.malfertheiner@infojuice.eu
§ The Penthouse, Carolina Court, Giuseppe Cali Street Ta' Xbiex, XBX1425, Malta
E-mail: pierluigi.casale@faculty.opit.com

Sentiment analysis (or opinion mining) identifies people's opinions, sentiments, attitudes, etc., towards entities such as consumer products, services and brands. This task has become a topic of increasing interest. Its applications have spread to almost every possible domain, including healthcare, financial services and political elections.

There are three main fields of application of sentiment analysis. **Brand reputation** involves sentiment analysis to help companies understand their reputation, credibility in the market and image perception. **Competitive intelligence** uses sentiment analysis to uncover and understand competitors' positions in the market. **Viral tracking** employs sentiment analysis to identify and monitor viral marketing campaigns through the tracking of public opinion.

Sentiment analysis is typically conducted using three main approaches: machine learning (ML), lexicon-based and hybrid approaches. **Machine learning** is the most widely used approach to sentiment analysis (Cui et al. 2023), leveraging advanced algorithms, such as support vector machines, neural networks, and transformers, in combination with linguistic features to perform sentiment classification. The **lexicon-based approach** uses lists of words that are commonly used to express positive or negative sentiments to identify the sentiment of the overall text. **Hybrid approaches** combine machine learning and lexicon-based approaches to improve sentiment analysis performance.

Researchers have mainly studied sentiment analysis at three levels of granularity: document-, sentence- and aspect-level. **Document-level** consists of establishing the overall opinion of a whole document as positive, negative or neutral (Pang and Lee 2008; Li and Li 2013). For example, given a product review, the review is classified based on the overall sentiment of the opinion holder about the product. The main assumption with this level of analysis is that each document expresses opinions on a single entity (e.g., a brand, product or service). Therefore, this analysis does not apply to documents that evaluate multiple entities. Both supervised and unsupervised learning approaches are used for document-level sentiment classification. **Sentence-level** determines whether a sentence expresses a positive, negative, or neutral opinion (McDonald et al. 2007). This level of analysis distinguishes objective sentences expressing factual information and subjective sentences expressing opinions (Wiebe, Bruce, and O'Hara 1999). Although sentence-level and document-level classification may use similar techniques, they differ significantly in their complexity and the type of analysis required. Sentences are smaller, self-contained units with clear syntactic structure, making it relatively straightforward to identify their sentiment. In contrast, documents are composed of multiple sentences that collectively form higher-level structures. This involves identifying relationships between ideas across paragraphs and analyzing overall discourse coherence. As a result, while the foundational techniques for both tasks might overlap, their application must be adapted to address the specific challenges of each level. As for document-level classification, sentence-level classification cannot be used in those applications that need to assign a sentiment score to a single target entity. **Aspect-level** focuses on the aspect or feature of entities (e.g., product features) (Pontiki et al. 2014; Basile et al. 2018), but it can also be applied to establish an opinion about a target entity as a whole. For instance, Hamborg and Donnay (2021) used this level of analysis to detect polar judgments toward target persons. As for document- and sentence-level classification, there are two main approaches for aspect-level classification: supervised learning approaches and unsupervised lexicon-based approaches. However, these approaches are not the same as their counterparts at document- and sentence-level, because aspect-level classification requires the target entity of an opinion to be identified.

Research into sentiment analysis has largely been conducted on product or service reviews, and movie reviews, where opinions are expressed openly, e.g.

> The website is very clear in all aspects of your services. Signing up for a new contract is extremely easy and you don't need to send anything by mail or email. Thank you very much.

Sentiment analysis on news articles has received far less attention. This task is more difficult compared to sentiment analysis on customer reviews. This happens because journalists often express their opinions indirectly, for example by highlighting some facts, while omitting others, or through the careful use of words. For example, the excerpt from the article below conveys a subtle positive sentiment about the mentioned company by highlighting its continued investment, expansion, strategic location selection, and positive contextualization within a larger trend of economic growth in the region. This approach avoids being explicitly promotional while still suggesting positive qualities about the company.

> Despite the difficulties posed by the COVID-19 pandemic, the company is still investing in Veneto. It has recently opened three new energy corners, bringing the total number of energy corners opened in 2022 to five. The company has selected these shopping centres for their convenience and strategic geographical position, creating new job opportunities in the process. This is part of a larger trend of companies investing in the region.

In this work, we use a case study of a company specialized in the analysis of brand reputation to evaluate machine learning models for sentiment analysis on news articles.

A team of experts annotates the sentiment expressed in local and national newspapers toward companies and public organizations that are subject to monitoring. Since the manual annotation of these articles is time-consuming and expensive, this study explores the use of machine learning to support its experts in their annotation tasks.

We tested several models for sentiment analysis, including traditional machine learning models like k-nearest neighbors (KNN), and transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT), Large Language Model Meta AI (LLaMA), and Generative Pre-trained Transformer (GPT). The implemented models were evaluated on a dataset of Italian, German and Ladin[1] news articles collected over the years and manually annotated with their sentiment polarity.

The results of these models were compared with the results of two baselines: (i) a lexicon-based method that uses a dictionary of sentiment words with associated sentiment polarity to compute a sentiment score for each article (ii) a frequency-based baseline that predicts the most frequent sentiment class within the dataset. In this comparison, the machine learning models outperformed the considered baselines. Furthermore, the results are also higher than those of Bastan et al. (2020) who evaluated BERT in a similar but not identical task. However, these findings reinforce the inherent challenges associated with this complex task.

1 A Romance language spoken in the Italian Alps

These study findings can be used by researchers in natural language processing as the baselines to compare their models on news articles. The implemented models can be used by the company to assist its annotators, resulting in time and cost savings.

2. Related Work

An exhaustive state of the art on the applications and challenges in sentiment analysis was presented by Pang and Lee (2008), Li and Li (2013), Birjali, Kasri, and Beni Hssane (2021), Wnkhade, Rao, and Kulkarni (2022) and Tan, Lee, and Lim (2023).

Practical implementations have been evaluated in various campaigns such as SemEval-2016 Task 4 (Nakov et al. 2016), SemEval-2017 Task 4 (Rosenthal, Farra, and Nakov 2017) and SemEval-2022 Task 10 (Barnes et al. 2022). Additionally, the SENTIment POLarity Classification Task in Evalita 2014 (Basile et al. 2014) and 2016 (Barbieri et al. 2016), as well as the Aspect-based Sentiment Analysis Task at Evalita 2018 (Basile et al. 2018) and Evalita 2020 (De Mattei et al. 2020), have provided valuable insights.

Bonadiman et al. (2017) further contribute by examining the real-world application of sentiment analysis models in industrial settings.

Different machine learning and lexicon-based approaches for sentiment classification have been applied to domains such as customer reviews and social media content.

2.1 Methods

Machine Learning methods for text categorization can also be applied to sentiment classification. Early studies were based on traditional machine learning models such as KNN (Huq, Ali, and Rahman 2017), Naïve Bayes (Kang, Yoo, and Han 2012) and SVM (Li and Li 2013; Rana and Singh 2016; Al Amrani, Lazaar, and El Kadiri 2018). More recently several deep learning approaches for sentiment analysis have been studied (Dang, Moreno-García, and De la Prieta 2020; Yadav and Vishwakarma 2019). These models do not require handcrafted features, because they can derive features directly from raw data. Even more recent studies are based on pre-trained language models such as BERT (Devlin et al. 2019) and its derivative models (Song et al. 2019; Sun, Huang, and Qiu 2019; Zeng et al. 2019). Lexicon-based uses a dictionary of words annotated with their sentiment score to classify documents as positive or negative (Jurek, Mulvenna, and Bi 2015). A sentiment score can be for example a simple polarity value +1 (positive) or -1 (negative). The final orientation of a document is obtained by calculating the semantic orientation values of the words in the document.

2.2 Language and Application Domain

Most studies on sentiment analysis have focused on domains such as movie reviews (Pang and Lee 2008; Socher et al. 2011; Singh et al. 2013), web discourse (web forums, newsgroups and blogs) (Abbasi, Chen, and Salem 2008; Pak and Paroubek 2010; Nakov et al. 2016; Zhang et al. 2020) and product reviews (Turney 2002; Dave, Lawrence, and Pennock 2003; Fang and Zhan 2015) in which authors tend to express their opinions explicitly.

Sentiment analysis of news articles has been extensively studied in various works. Godbole, Srinivasaiah, and Skiena (2007) developed a scalable system for processing large corpora of news articles and blogs. The system constructs sentiment lexicons by expanding small seed lists using WordNet-based path analysis, filtering ambiguous terms and refining results manually. Sentiment is identified by marking relevant terms,

adjusting polarity for negations and modifiers, and linking sentiment to entities via co-occurrence in sentences. Balahur et al. (2010) used lexicons like SentiWordNet and in-house resources to compute sentiment scores within word windows around target mentions while excluding category-defining words for improved accuracy. The focus was on analyzing quotations (reported speech) in news articles. Hamborg and Donnay (2021) introduced a dataset for target-dependent sentiment classification (TSC) in political news. Their neural model combines a pre-trained language model (e.g., RoBERTa) and external knowledge sources (e.g., sentiment and psychometric dictionaries). Bastan et al. (2020) annotated a dataset of news articles with the sentiment expressed toward some target persons in the articles. This dataset was used to evaluate machine learning models, demonstrating the complexity of sentiment classification in the news domain. Our work was inspired by this study with three key differences: (i) they conducted their experiments with articles that focus on one single named entity of interest. On the other hand, an article in our dataset can cover multiple named entities; (ii) the objective of their work was to identify the sentiment expressed towards persons, whereas our task focuses on the sentiment towards organizations; (iii) they only considered news articles written in English for which several annotated resources and annotation tools exist, while our study focused on articles in Italian, German and Ladin, for which, compared to English, fewer resources are available.

Cross-domain sentiment analysis was studied by Li et al. (2017) and Du et al. (2020). These studies show that sentiment classification is highly sensitive to the domain from which the training data are extracted. Indeed, a classifier trained on one domain often performs poorly on another domain since sentiment in different domains can be expressed in different ways.

Cross-language sentiment classification is the sentiment classification of opinion documents in multiple languages. Among the published studies, it is worth mentioning the study by Zhou, Wan, and Xiao (2016).

Researchers have expanded sentiment analysis into diverse domains, including literature and historical texts. For instance, Sprugnoli et al. (2023) made a study on sentiment analysis in Latin poetry. This study highlights the adaptability of sentiment analysis techniques across different textual genres and historical eras.

3. Sentiment Analysis

Sentiment analysis, or opinion mining, is a field of natural language processing (NLP) and computational linguistics that focuses on identifying and extracting subjective information from text. It aims to analyze and understand the emotional tone, opinions, attitudes, and beliefs expressed in text and other forms of human communication. This study explores the task of identifying the overall opinion expressed by the author of a news article toward a target entity mentioned in the article. According to Liu (2015), an opinion can be defined as follows:

an opinion is a quintuple (e, a, s, h, t), where e is the target entity, a is the aspect of entity e about which the opinion has been given, s is the sentiment of the opinion about aspect a of entity e, h is the opinion holder, and t is the opinion posting time; here, s can be positive, negative, or neutral

Sentiment analysis based on this definition is called aspect-level sentiment analysis. Not all applications for sentiment analysis need to be defined as a quintuple. For

example, in sentiment analysis for brand reputation, the interest is only in opinions about a target brand as a whole. This approach is called entity-level sentiment analysis.

Consider the following scenario in which a document and a target entity (ORG) are provided.[2] In this case, the objective is to discern the sentiment expressed in the document about this entity, which in this instance is positive.

> [ORG] won first place in the [EVENT] ranking in [LOC] category. The study, conducted by the German Institute for Quality and Finance (ITQF), evaluated the level of customer satisfaction in terms of service quality, efficiency, and timeliness of responses. This remarkable achievement is a testament to [ORG]'s unwavering commitment to excellence and their dedication to providing exceptional customer experiences.

To assess the quality of the classified documents, it is required to use standardized evaluation scores. A frequently used error measure is the F_1 score, which is a combination of Precision and Recall (Rijsbergen 1979). This measure is defined as follows: Precision takes the number of texts that are correctly classified for a given sentiment tag and divides it by the number of texts that were predicted (correctly and incorrectly) as belonging to that tag. Recall takes the number of texts that are correctly classified for a given tag and divides it by the number of texts that should have been predicted as belonging to that tag. F_1-micro score is the harmonic mean of these two measures.

$$F_1\text{-micro} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Since F_1-micro is dominated by the classifier's performance on more frequent categories, F_1-macro (defined as the average F_1-micro across all classes) is often preferred to F_1-micro on unbalanced datasets like the dataset used in this study.

$$F_1\text{-macro} = \frac{1}{N} \sum_{i=1}^{N} F_1\text{-micro}(i)$$

where N is the number of classes.

4. Use Case

In this study, we present the case of an Italian company specializing in Media Intelligence, namely Infojuice Srl.[3], and referred to as the "Company" hereafter in this article. It provides clients with personalized media intelligence services like press reviews, media analysis, and PR insights. Clients are distinguished by very heterogeneous fields of interest. Depending on their activities, interest can range from the financial sector, banking, insurance to cultural and art contexts. The client portfolio also includes public institutions (local, national and international) as well as private companies. The Company ensures compliance with copyright laws and has held membership in

2 The names of the target organization ("ORG"), event ("EVENT") and location ("LOC") have been altered to preserve the anonymity of the organization mentioned in the document.

3 https://www.infojuice.eu/

FIBEP (International Federation of Press Clipping Services) since 2012, alongside being a founding member of FederRassegne, Italy's premier association for Media Intelligence Organisations. Among its services, the Company monitors sentiment expressed in local and national printed newspapers towards its clients.

Currently, sentiment extraction from news articles is conducted manually by a team of expert annotators. On average, 2 minutes are required to read and annotate a single news article, while there are as many as 200 to 300 articles to review each day. Since most of these articles do not contain a clear opinion or sentiment concerning the target entity, a significant proportion of the effort expended on this activity is unproductive. To overcome this issue and assist the team in their annotation process, this study explores the use of machine learning for sentiment classification. We conducted an extensive study to train and evaluate various machine-learning models on a dataset of 8,978 annotated news articles annotated in 2021.

5. Dataset

Conventional sentiment analysis datasets for research typically aim for a balanced distribution of positive, negative, and neutral examples. When not perfectly balanced, the examples generally remain within the same order of magnitude across sentiment categories. For instance, the Internet Movie Database (IMDb) (Maas et al. 2011) dataset ensures a strictly balanced setup with an equal number of positive and negative reviews, omitting neutral examples. Similarly, the Sentiment140 dataset (Go, Bhayani, and Huang 2009) maintains an equal distribution of positive and negative sentiment classes, facilitating fair sentiment analysis tasks. The SemEval-2017 Task 4 Subtask A dataset (Rosenthal, Farra, and Nakov 2017), although not strictly balanced, contains examples across sentiment categories that are of the same order of magnitude. However, the datasets of news articles used in this study show an uneven distribution, with a prevalence of neutral articles compared to the other two categories. This is because news articles typically adopt a neutral or unbiased tone, especially when reporting on established institutions or entities. This tendency stems from journalistic efforts to maintain objectivity and provide balanced coverage, avoiding excessively positive or negative assessments.

The dataset under consideration includes 8,978 news articles published in Italian, German, and Ladin languages spanning twelve months (January-December 2021). The articles are taken from Italian national and regional newspapers. For example, "Il Corriere della Sera" for national newspapers and "Alto Adige" for regional newspapers. Each article is associated with one of ten entities monitored by the brand monitoring services of the Company. The dataset includes entities such as universities, industry associations, museums, banks and energy providers, as shown in Table 1.

Notably, the dataset exhibits a significant imbalance, containing 6,695 neutral articles, 2,142 positive articles, and only 141 negative articles.

Distinguishing itself further from standard research datasets, the dataset of this study supports sentiment annotations based on contextual and external information, not just on the information in the article. Such annotation method was also employed by Bastan et al. (2020). These annotations were created by expert annotators who adhered to the following guidelines:

- to judge the sentiment orientation of an article the annotators should respond to this question: if you were in the position of the entity mentioned in the article, would you experience contentment?

Table 1
Distribution of entities across different sectors.

Entity type	Occurrences
local authority	2,910
university	1,603
industry association	1,315
museum	1,013
bank	727
energy provider	651
insurance provider	526
multinational service	189
sectoral association	44

Table 2
Statistics on positive, neutral, and negative examples, followed by annotated documents and tokens for Italian, German, and Ladin (based on space delimiter).

		Italian	*German*	*Ladin*	*All*
Sentiment	Positive	1,277	798	67	2,142
	Negative	95	45	1	141
	Neutral	4,563	2,088	44	6,695
Metrics	Documents	5,935	2,931	112	8,978
	Tokens	3,053,268	1,284,718	80,897	4,418,883

- each article must be annotated with its sentiment orientation (positive, negative or neutral) towards the target entity of the article

The annotators are seasoned professionals employed by the Company, each with over 15 years of experience in sentiment analysis and brand monitoring. While no formal inter-annotator agreement metric is used to evaluate annotation quality, the annotation process ensures reliability and consistency through a well-structured workflow. This includes the implementation of the Specialized Onboarding Process for Annotation, designed to align annotators with the specific requirements of each client. The process consists of the following steps:

1. Initial Client Consultation: A detailed discussion is held with the client to understand their specific context, including any particular elements or details that require attention.
2. Preparation and Approval of a Brief: Based on the consultation, a brief is drafted outlining the annotation guidelines and objectives. This document is shared with the client for review and approval.
3. Brief Sharing with the Annotator: Once approved, the brief is provided to the annotator responsible for the project.

4. Initial Annotation Review: Before starting the main annotation task, an initial review session is conducted between the annotator and the department manager to align on expectations and ensure clarity regarding the guidelines.
5. Independent Annotation: The annotator proceeds autonomously, handling the sentiment analysis of the articles. However, they may consult the manager for clarification on specific cases or incorporate feedback when necessary.

Sentiment analysis for news articles differs from that for product and service reviews, or in posts on Twitter and Facebook, in the following key aspects.

Implicit Sentiment Analysis: Instead of directly expressing sentiment through emotional language, news articles often convey it indirectly through the careful selection of words or simply by reporting some facts, while omitting others. Below, we report an example of an article extracted from the dataset, which expresses a positive sentiment toward an Italian University (denoted as "ORG" to protect privacy). This article does not explicitly praise the university's initiative. Instead, it highlights the practical benefits of the vaccination offer, such as convenience and accessibility.

> [ORG], from [TIME] vaccinations for everyone. The new academic year will start on [TIME] and [ORG] wants to offer students, professors and administrative staff the opportunity to get vaccinated without having to make an appointment. The offer is open to all citizens.

Granular Sentiment Analysis: Unlike reviews or social media posts, which typically revolve around a single entity, news articles usually contain multiple entities. This requires sentiment analysis at the entity level rather than the document level, adding complexity to the task. For example, in the following excerpt, the sentiment analysis is positive for the anonymized entities ORG#1 and ORG#3, while it is negative for the entity ORG#2.

> [ORG#1] wins and the contract for the catering service in a part of the school canteens in [LOC] remains valid. In [TIME], as you will remember, fierce controversies erupted over the quality of the food served to children. [ORG#1] had then decided not to renew the contract with [ORG#2] and the service (for an estimated annual amount of 6 million) had been awarded to [ORG#3]. Recourse to the Tar immediately started, based on alleged errors contained in the offer of the winning company, while [ORG#3] also presented a recourse against alleged errors in the offer of [ORG#2] (second in the race). A legal mess from which the Tar now comes out with a judgment with which the correctness of [ORG#1] (and of the awarding commission) is recognized in the whole affair «and the requests aimed at obtaining the declaration of the ineffectiveness of the contract and the sub-entry in the same must be consequently rejected».

Sentiment Spillover: News articles often address entities that play significant roles in society and maintain connections to other individuals or organizations. These relationships imply that opinions articulated about one entity can indirectly influence

the public perception of others. For instance, in the following passage, the negative things that the journalist is reporting about a person (PER#1), who is the president of the mentioned organization, are also affecting the way people think about the organization (ORG) itself.

> [PER#1] sues [PER#2] but loses and is sentenced to pay 21,000 euros. In the end, after suing provincial councilor [PER#2] for defamation, it will be him, [PER#1], a politician and president of [ORG], who will have to compensate [PER#2] with 15,000 euros plus legal fees for a total of 21,000 euros. This was decided by the Court of [LOC] which issued the first-degree judgment, rejecting the compensation request made by [PER#1] against [PER#2] and instead condemning the MP to compensate the provincial councilor for his statements.

Context-Aware Sentiment Analysis: Compared to reviews and social media posts, news articles tend to be significantly longer. This makes it difficult to identify the writer's attitude towards a target entity of interest. The provided passage presents a subtle analysis of the company (referred to as ORG), with the author's sentiment only revealed in the conclusion. To fully comprehend the author's viewpoint, readers must traverse the entire narrative of PER#1.

> «I lost a battle, not the war». This is what [PER#1], an innkeeper and environmentalist, said after the ruling by the Council of State that ruled in favor of the owner of the malga. The road that leads from [LOC] to the malga can also be built in the last stretch of about one kilometer. [PER#1], together with other environmentalists, had been the protagonist, years ago, of a battle against the road. Now he talks about the ruling that sees him defeated. But it seems to understand that he has no intention of giving up his fight against the completion of the road. Then the stab at [ORG] and the councilor [PER#2]: «They have shown, once again, that they are insensitive to certain problems».

5.1 Data Availability

While the sentiment analysis dataset employed in this study cannot be distributed due to privacy concerns, we have provided a comprehensive description of the dataset, including entity types, news article sources, annotation guidelines, preprocessing procedures, model configurations, and experimental results. This should enable researchers to replicate our findings and develop their sentiment analysis models using similar data sources and methodologies.

6. Method

We wanted to establish the usefulness of the dataset of the Company to train a classifier for the task of sentiment analysis on news articles. To do this, we split the dataset into training (50% of data), development (25%), and test datasets (25%) as shown in Table 3.

KNN is set as the baseline to compare state-of-the-art pre-trained models to a traditional machine learning model. All the evaluated models were trained on the training dataset and tuned on the development dataset. The resulting best configuration on the development dataset was tested on the test dataset. The models and baselines were evaluated by F_1-macro as well as F_1-micro. Additionally, we calculated F_1-macro

Table 3
The dataset is split into train, dev, and test datasets.

	Positive	*Negative*	*Neutral*
Train	1,038	83	3,368
Dev	549	22	1,673
Test	555	36	1,654

on positive and neutral examples (namely F_1-macro-pos-neut) to evaluate the models' performance on labels with high representation in the dataset.

The experiments presented in this paper were performed with a Linux PC.[4]

6.1 Models Evaluation

KNN is a well-known ML method that has been widely used for text classification (Huq, Ali, and Rahman 2017). In this study, the Scikit-learn library[5] was used to implement KNN in Python 3.9.12. KNN was chosen as a comparison baseline because it is a well-established, traditional machine learning algorithm frequently used in text classification tasks. The algorithm works by comparing a given test instance (i.e. a document or text) to the training instances in the feature space. Each instance is represented as a vector of features, typically derived from the text (e.g., through techniques such as TF-IDF or word embeddings). The KNN algorithm identifies the k nearest neighbors to the test instance based on a chosen similarity metric, usually Euclidean distance or cosine similarity, which measures the proximity of feature vectors. Once the nearest neighbors are found, the algorithm assigns the label that is most frequent among these neighbors to the test instance. We represented the annotated articles as vectors of weighted features (or terms) computed with TF-IDF. Specifically, the TF-IDF score was calculated for each news article by concatenating the title and text body. The parameters of KNN were set to their default values, i.e. 5 for K value and Euclidean distance for document similarity measure.

Traditional machine learning models like KNN may require a large amount of data to achieve high performance. Unfortunately, available annotated datasets for sentiment analysis, including our dataset, only consist of a few thousand labeled documents. Given this limitation, we decided to include pre-trained models like BERT (bert-base-uncased model) in our comparison, as they can often achieve better results than traditional machine learning models even on small datasets. The English BERT model was included mainly as a benchmark to evaluate its performance. Due to its popularity and proven success in many NLP tasks, we believe BERT can serve as a valuable reference point, even in contexts where the dataset is predominantly non-English. Additionally, BERT-ITA[6], a version of BERT specifically trained on Italian text, was tested to assess whether a language-specific model could achieve better performance compared to a multilingual one like XLM-RoBERTa. To conduct these experiments, we used Simple-

4 Ubuntu 20.04, GeForce RTX 2080 ti
5 `https://scikit-learn.org`, version 1.0.2
6 dbmdz/bert-base-italian-uncased

Transformers v0.63.6[7], which is a deep learning library built over the popular HuggingFace.[8] Since each article in our dataset consists of many attributes (e.g., article title and text), we evaluated different ways of representing an article for sentiment classification.

In our study, XLM-RoBERTa (xlm-roberta-base model) (Conneau et al. 2020) was tested to take advantage of the annotation of the multilingual corpus. This is possible because multilingual models such as XLM-RoBERTa are pre-trained on corpora in multiple languages (including Italian and German) and hence they can be used for text classification tasks in more than one language.

One of the main hyper-parameters that may affect the accuracy of pre-trained models is the number of learning epochs. While too many epochs can lead to the overfitting of the training dataset, too few epochs may result in an underfitting model. The experiments described above were repeated varying the number of epochs between 1 to 10. As regards the other main hyper-parameters, the batch size was set to 8, max sequence length and learning rate were set to 512 and 2e-5, respectively.

We compared the results of our models with those of two baseline methods often used in the literature. A first baseline was produced by a lexicon-based approach, using Sentix[9], an Italian sentiment lexicon, to calculate sentiment scores for each document. Each word in Sentix is disambiguated with its part-of-speech tag and associated with its positive and negative scores. To compute this baseline, a document is first tokenized and tagged with the spaCy part-of-speech tagger.[10] Then, the tokens and their corresponding part-of-speech were matched with entries in the lexicon to determine their sentiment scores. Finally, the overall sentiment of the document is obtained by calculating the sum of the sentiment score of its words. This baseline is specifically designed for the Italian language, exploiting a lexical resource tailored to it. For the German and Ladin texts, we assigned the most frequent sentiment class in the dataset (neutral) as a pragmatic solution to ensure consistency across the languages. However, this approach introduces an imbalance, as the sentiment analysis for Italian is more detailed and adapted to its linguistic features, while the German and Ladin texts lack the same level of refinement. The second baseline predicts (i) the most frequent sentiment class (neutral) within the dataset and (ii) the most frequent sentiment class for the target entity.

We used GPT-3.5-Turbo and Text-davinci-003 from OpenAI[11] to leverage their pre-trained contextual understanding to analyze the sentiment of news articles. These models were employed to identify the underlying sentiment, even when it was not explicitly expressed, as is the case in some articles within the dataset. Text-davinci-003, the newer and more advanced model, is noted by Ye et al. (2023) to outperform GPT-3.5-Turbo in tasks such as Machine Reading Comprehension, which may be due to the smaller model size of GPT-3.5-Turbo. Furthermore, Text-davinci-003 was trained on a more recent dataset and uses a distinct training strategy known as RLHF (reinforcement learning with human feedback). This approach has been shown to enhance the model's ability to follow instructions and generate more coherent text. GPT-3.5-turbo was designed to be a more efficient version of text-davinci-003. It has a smaller model size and uses a different training strategy that is optimized for chat tasks. To compare the performance of these two models, we used the Microsoft Azure API version 2023-05-15 and presented

7 https://simpletransformers.ai
8 https://huggingface.co
9 https://valeriobasile.github.io/twita/sentix.html
10 https://spacy.io/models/it, model lt_core_news_sm-3.7.0
11 https://platform.openai.com/docs/model-index-for-researchers

the same prompt to both models.[12] Moreover, we set the temperature parameter to 0 in order to ensure a certain degree of consistent outcomes.

In addition to the models previously mentioned, we also experimented with Llama, which is a family of large language models (LLMs) created by Meta AI. Released in multiple versions (Llama 1-3), these models are trained on massive datasets of text and code, enabling them to support a wide range of NLP tasks. According to the authors of the Llama models, the latest version, Llama 3, offers improved performance and functionalities compared to the previous models (MetaAI 2024). It is also committed to open-source accessibility. In our evaluation, Llama-3-8B was evaluated using the model implementation available on Hugging Face.[13] Llama outputs often included multiple labels, numeric values, or repeated text sequences influenced by its pre-training data. To address this, we experimented with various prompts to guide the model response. Ultimately, we found a prompt[14] that partially mitigated these undesired outputs. Furthermore, to ensure a more consistent evaluation process, we adopted a two-step approach. First, we assigned a unique sentiment label (positive, negative, or neutral) to responses with clear predictions that contained only one of the three labels (78.4% of the model predictions). Second, outputs that deviated from the expected format (21.6% of the model predictions) were assigned a *neutral* label.

6.2 Preprocessing

Thorough data preparation, accurate data representation, and balanced data distribution are crucial for achieving high accuracy and reliability in sentiment analysis.

Data Preparation and Representation: Training and test data must be converted to an appropriate format before feeding into machine learning models. Our models require input data to be in Pandas DataFrame format, which has to include two columns. One column contains the article text, while the other column contains the sentiment label associated with the text. The following 2 steps were applied for data preprocessing. First, we used a variation of the procedure by Bastan et al. (2020) to mask the target entity of an article with the "TGT" token. More precisely, in our procedure, we used string matching for entity replacement, while they used a more sophisticated approach for coreference resolution based on the analysis of the spaCy tool. Masking target entities serves two purposes: to prevent models from easily learning biases towards entities and to target the entity of interest. Second, we removed those fields that were not useful for the final classification (i.e. ArticleID, Journal, Language, Date, PublicationPage). Subsequently, we investigated two different approaches to represent articles for sentiment classification: using only the article's title as an input feature and combining the article's title and text into a single string as an input feature.

Data Balancing: An unbalanced dataset may contribute to decreasing the ability of the model to correctly classify the articles of the minority class. To overcome this issue, we experimented by balancing the training dataset with Italian and German customer reviews (Table 4). The Italian reviews were selected randomly from a dataset of 2,179,676 reviews of products and services, referred to as the ItalianReview dataset throughout this paper. This dataset was compiled from multiple sources, including

12 Provide a comprehensive sentiment analysis of the news article provided, identifying the overall sentiment (positive, negative or neutral) directed towards $TARGET_ENTITY

13 `https://huggingface.co/meta-llama/Meta-Llama-3-8B`

14 Express the overall sentiment of the news article towards the target $TARGET_ENTITY. Should the sentiment be classified as Positive, Negative, or Neutral?

Italian reviews from Amazon.it, Trustpilot[15] and Tripadvisor.[16] The German reviews were selected randomly from a public dataset of Amazon's reviews, referred to as the GermanReview dataset in this paper.[17] Both datasets express sentiment using a star rating system, with 5 stars indicating the highest level of positivity. To standardize this annotation with our dataset, we labeled reviews with 4 or 5 stars as positive, those with 3 stars as neutral, and those with 1 or 2 stars as negative. To make sure the Italian and German reviews reflected the language distribution in the training dataset, the number of Italian examples was set to be double the number of German examples. Futhermore, the sum of positive and negative examples was set to match the number of neutral examples in the training set (3,368), which represents the most populated category. Once balanced, this dataset was used to train the models.

Table 4
The balanced training dataset contains an equal number (3,368) of positive, negative and neutral articles.

	Positive	*Negative*	*Neutral*
the Company's train dataset	1,038	83	3,368
ItalianReview	1,553	2,191	0
GermanReview	777	1,094	0
ALL	3,368	3,368	3,368

Data Enrichment: The accuracy of ML models largely depends on the quantity of training data. In our study, the training data was enriched through two different methods. The first method consists in using data augmentation techniques to generate new training examples from existing data. To do that, we exploited state-of-the-art machine translation models like MarianMT[18] to translate 2,999 Italian articles of the training data into German and 1,439 German articles of the training data into Italian. This creates a training dataset twice the size of the original data. The second method increases the quantity of the training data by adding 120,000 Italian reviews and 60,000 German reviews selected from the ItalianReview and GermanReview datasets (Table 5), respectively. The two methods used for expanding the training dataset were finally evaluated by fine-tuning the XLM-RoBERTa model on the two enlarged training datasets.

Table 5
The training dataset was expanded by adding customer reviews and performing data augmentation through translation between Italian and German datasets.

		Positive	*Negative*	*Neutral*
customer reviews	Italian	40,000	40,000	40,000
	German	20,000	20,000	20,000
translated articles	Italian	363	25	1,051
	German	642	58	2,299

15 https://github.com/AlessandroGianfelici/italian_reviews_dataset
16 https://www.kaggle.com/datasets/alessandrolobello/italian-tripadvisor
17 https://huggingface.co/datasets/amazon_reviews_multi
18 https://huggingface.co/docs/transformers/model_doc/marian

7. Results

This section reports the experiments done to find the best system configuration on the development set. The resulting best configuration was tested on the test dataset. As reported in Section 6, the pre-trained deep learning models were set up with the same configuration, i.e. batch size was set to 8, max sequence length and learning rate were set to 512 and 2e-5, respectively. The number of epochs varies for each experiment, while each article in the dataset was represented by concatenating its title and text.

As presented in Table 6, all machine learning models showed higher performance compared to the baselines. Then, multilingual RoBERTa (XLM-RoBERTa) outperforms the other evaluated models. Notably, KNN performed better than BERT, which is a model pre-trained on English data. Interestingly, XLM-RoBERTa (F_1-macro: 64.19) outperformed BERT-ITA (F_1-macro: 57.16) on the subset of Italian documents from the development set, while KNN (F_1-macro: 56.24) showed similar performance to BERT-ITA (F_1-macro: 57.16). The lexicon-based baseline (F_1-macro: 34.02), which uses a lexical resource for Italian and assigns the most frequent tag (neutral) for German and Ladin, performed better than the baseline based solely on the most frequent tag (F_1-macro: 28.47).

Table 6
F_1 measures of the machine learning models and baselines (lexicon-based, most-frequent-tag-per-entity, most-frequent-tag) trained on the training dataset and evaluated on the development dataset.

	F_1-macro	*F_1-micro*	*F_1-macro-pos-neut*	*epochs*
XLM-RoBERTa	**65.22**	83.38	78.68	6
KNN	57.00	78.88	70.50	-
BERT-ITA	53.67	77.94	69.98	5
BERT	48.40	79.32	72.60	7
most-frequent-tag-per-entity	48.00	79.00	71.50	-
lexicon-based	34.02	69.79	51.04	-
most-frequent-tag	28.47	74.55	42.71	-

Table 7 shows that accuracy is higher for categories that are more common in the training dataset.

Table 7
Precision, Recall and F_1-micro measures of XLM-RoBERTa computed on the positive, negative and neutral articles of the development dataset.

	Precision	*Recall*	*F_1-micro*	*support*
Neutral	89.03	88.82	**88.93**	1,673
Positive	68.36	68.49	68.43	549
Negative	36.00	40.91	38.30	22

In Table 8, we observe that fine-tuning the XLM-RoBERTa model on the training dataset balanced with Italian and German reviews (F_1-macro: 63.39) does not improve the results compared to using the original training data (F_1-macro: 65.22). Extending

the training data by translating Italian articles into German and vice versa (F_1-macro: 63.40) or concatenating the training data and customer reviews (F_1-macro: 60.01) performs poorly than using the original training data. Even though the ItalianReview and GermanReview datasets contain a substantial amount of customer reviews, fine-tuning the model solely using these reviews leads to a substantial decrease in performance (F_1-macro: 25.00). Additionally, fine-tuning the XLM-RoBERTa model on only the sentences mentioning the target entities (F_1-macro: 62.67) is not as good as fine-tuning the model on the whole document.

Table 8
F_1 measures produced by fine-tuning XLM-RoBERTa on the training dataset, the balanced train dataset, the training dataset enriched with customer reviews or data translation, the customer reviews only, or the sentences in the training dataset mentioning the target entities. All the tested configurations were evaluated on the development dataset.

training data	F_1-*macro*	F_1-*micro*	F_1-*macro-pos-neut*	*epochs*
This Work's Train Dataset	**65.22**	83.38	78.68	6
enriched with article translations	63.40	82.44	77.32	6
balanced	63.39	83.11	78.42	7
enriched with customer reviews	60.01	82.01	77.50	6
sentences	62.67	82.17	75.40	7
customer reviews	25.00	29.00	35.00	6

All the experiments reported above were done by representing the articles of the dataset with their title and text attributes. We also explored an alternative approach by representing articles solely based on their titles. In this comparison, the model that employed both title and text achieved higher F1-macro scores (F_1-macro: 65.22) compared to the model that used only titles (F_1-macro: 41.27).

We employed the most effective model configuration, derived from the development dataset (XLM-RoBERTa) to annotate the test dataset (Table 9). The untuned GPT and Llama models achieved significantly lower F1-macro scores (F_1-macro: 49.00 and F_1-macro: 27.00) compared to the optimized XLM-RoBERTa model (F_1-macro: 58.32). Notably, our model's performance is higher than the results reported by Bastan et al. (2020) who tested BERT on a dataset of English news articles, achieving an F_1-macro score of 48.00.

To ensure a robust evaluation of model performance, particularly given the dataset's limited size, we also performed an experiment using cross-validation with XLM-RoBERTa, which provided mean (F_1-macro: 60.50) and standard deviation values (4.48). Additionally, we performed bootstrapping on the test set, obtaining an average F_1-macro score of 58.13 with a 95% confidence interval of (52.99, 63.59).

The confusion matrix presented in Table 10 calculated for XLM-RoBERTa on the test set shows that there are only a few cases of positive examples being misclassified as negative and negative examples being misclassified as positive.

XLM-RoBERTa performs equally for German and Ladin, according to Table 11. Italian produces significantly higher results than German and Ladin. However, when looking at the F_1-macro-pos-neut score calculated on the positive and neutral categories, we see that for these two most represented categories in the dataset, the measured accuracy values are in line between the three languages.

Table 9
Comparison of Model Performance on Test Dataset: Evaluating the Performance of XLM-RoBERTa after fine-tuning on training and development datasets against untuned GPT and Llama models.

	F_1-*macro*	F_1-*micro*	F_1-*macro-pos-neut*
XLM-RoBERTa	58.32	82.67	77.28
Text-davinci-003	49.00	69.00	61.50
GPT-3.5-turbo	43.00	62.00	57.00
Llama-3-8B	27.00	42.00	38.50

Table 10
Confusion matrix of the base-model on the test set.

Gold/Prediction	*Neutral*	*Negative*	*Positive*
Neutral	1511	7	136
Negative	31	5	0
Positive	214	1	340

Table 11
Detailed F_1 measures for the three languages using XLM-RoBERTa.

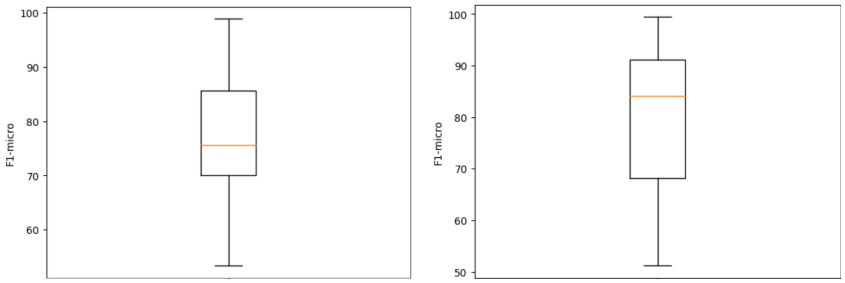
	F_1-*macro*	F_1-*micro*	F_1-*macro-pos-neut*
Italian	61.13	84.31	77.41
German	54.07	79.65	76.10
Ladin	52.96	96.96	78.57

Figure 1 highlights that the F_1-micro scores across the 10 entities in the dataset are much closer to each other compared to the F_1-micro scores for the Neutral category alone. The majority of the data points for the F_1-micro scores in the Neutral category are concentrated towards the upper segment of the boxplot, with a longer tail extending downwards.

To compare the difficulty of sentiment analysis on news articles to that on customer reviews, we evaluated the performance of XLM-RoBERTA on a subset of customer reviews selected from the ItalianReview and GermanReview datasets. This subset was randomly chosen to mirror the category distribution found in the Company's dataset. On these selected reviews, XLM-RoBERTA achieved an F_1-macro score of 64.0 and F_1-macro-pos-neut score of 95.5. The low F_1-macro values compared to those of F_1-macro-pos-neut are justified by the tendency of the classifier to avoid predicting the underrepresented negative category in the training data.

8. Discussion

News articles pose a significant challenge for sentiment analysis due to their inherent linguistic complexity. They often contain complex language and are less likely to contain



(a) (b)

Figure 1
On the left, the boxplot shows the F_1-micro scores for all three categories: Positive, Negative, and Neutral, across the 10 entities in the dataset. On the right, the boxplot shows F_1-micro scores solely for the Neutral category.

direct expressions of sentiment. Despite this, the overall outcome of the experiments shows that the dataset used in this study can be used to train a machine learning model for sentiment classification, which can assist annotators in their annotation task.

Regarding the effectiveness of the evaluated models, Table 6 compares the results of machine learning models like XLM-RoBERTa (F_1-macro: 65.22) with the result of the lexicon-based approach (F_1-macro: 34.02) and the most frequent tag baselines (F_1-macro: 28.47) on the development dataset. Significantly, the machine learning models outperformed the baselines considered. The poor performance of the lexicon-based baseline can likely be attributed to its simplicity, which restricts its ability to address contextual variations. Specifically, the implemented system has the limitation of not considering the context in which words appear, leading to potential inaccuracies when interpreting words with multiple meanings, or negations, which can significantly affect the sentiment analysis results.

XLM-RoBERTa (F_1-macro: 65.22) outperforms the other evaluated models, such as BERT-ITA (F_1-macro: 53.67), and even when compared to English BERT (F_1-macro: 48.40), it performs better on the entire development dataset. This can be explained by the fact that XLM-RoBERTa, being a multilingual model, is well-suited for handling all three languages in the dataset, while BERT-ITA is specifically designed for the Italian language, and English BERT is tailored for English, which is not present in the dataset used in this study. One possible explanation why XLM-RoBERTa (F_1-macro: 64.19) performs better than BERT-ITA (F_1-macro: 57.16) even on a subset of Italian news articles is that XLM-RoBERTa trained on a diverse range of languages, can capture linguistic patterns and structures common to them. In comparison, models like BERT-ITA are pretrained on a narrower corpus. Moreover, the performance of KNN (F_1-macro: 56.24), which aligns closely with that of BERT-ITA (F_1-macro: 57.16) on the subset of Italian news articles, suggests that simpler models like KNN can effectively capture domain-specific patterns, particularly when the training set includes representative examples that help annotate the test data.

Our dataset is highly unbalanced and this causes a bias of the classifier towards the majority class in the dataset (Table 7), with an F_1-micro of 88.93 for the neutral sentiment class compared to 38.30 for the negative sentiment class. In an effort to address this concern, we balanced the training data by concatenating the training data with customer reviews (Table 8). Our results seem to confirm the observations of Li et al. (2017) and Du

et al. (2020) that a classifier trained on one domain often performs poorly on another domain since sentiment in different domains can be expressed in different ways. In our view, such observations may also explain why relying solely on substantial amounts of customer reviews to train the classifier does not lead to a corresponding increase in accuracy (Table 8).

Turning now to how the evaluated models perform on the test dataset, the results in Table 9 show that XLM-RoBERTa fine-tuned on the dataset used in this study (F_1-macro: 58.32) performs significantly better than BERT tested by Bastan et al. (2020) on another dataset consisting of news articles (F_1-macro: 48.00). Moreover, the model processes and annotates up to 102 newspapers per second on a PC with Ubuntu 20.04 and a GeForce RTX 2080 Ti, exceeding the Company's daily workload (200 to 300 news articles per day).

The F_1-macro scores from cross-validation (60.50) and bootstrapping (58.13, with a 95% confidence interval of 52.99–63.59) are close to the test dataset score of 58.32. This alignment across different evaluation methods and test partitions confirms that the model produces reliable results, demonstrating its ability to generalize well across varying data splits.

We used GPT-3.5-Turbo, Text-davinci-003 and Llama models to try to leverage their vast knowledge base and understand the nuances of language used by journalists to express their opinions. Table 9 shows that XLM-RoBERTa, fine-tuned on our training dataset, performs significantly better than GPT-3.5-Turbo (F_1-macro: 43.00) and Text-davinci-003 (F_1-macro: 49.00).

Llama-3-8B's performance (F_1-macro: 27-00) was difficult to evaluate because its outputs were inconsistent. It sometimes generated multiple labels, numbers, or text snippets probably derived from its pre-training data. To address this, we used prompts to get it to give a single sentiment label and assigned a *neutral* label for unexpected outputs. Few-shot prompting offers a potential future exploration for evaluating Llama-3-8B. However, this technique requires careful selection of representative examples from the dataset. This is particularly important because the sentiment studied in this work is multifaceted. It can be both explicit and implicit, and it varies across languages and the domains of the named entities mentioned in the text.

The performance of XLM-RoBERTa and Text-davinci-003 was evaluated in different sentiment annotation scenarios in our dataset (see Section 5). Implicit Sentiment Analysis: In the document in which the sentiment is expressed indirectly, XLM-RoBERTa effectively detected the positive sentiment about a university's vaccination initiative. It likely identified key phrases such as "vaccinations for everyone", "open to all citizens", and "opportunity", which emphasized inclusivity and accessibility. Text-davinci-003, which leverages external knowledge, probably recognized the positive sentiment by associating the initiative with public health support. Granular Sentiment Analysis: In the document in which the sentiment varied between entities, XLM-RoBERTa accurately identified the sentiment toward ORG#1 as positive, likely recognizing the pattern "ORG#1 wins". Text-davinci-003 performed similarly to XLM-RoBERTa, obtaining similarly accurate results, likely by leveraging the same patterns as XLM-RoBERTa. Sentiment Spillover: In the article where negative sentiment toward an individual (PER#1) influenced the sentiment associated with their organization (ORG), XLM-RoBERTa labeled ORG as neutral, highlighting its limitations in identifying indirect sentiment. In contrast, Text-davinci-003 demonstrated a deeper understanding of how negative perceptions of a person can affect the reputation of their associated organization, accurately labeling ORG's sentiment as negative. Context-Aware Sentiment Analysis: In a longer article where sentiment was implied in the conclusion, XLM-RoBERTa failed to identify

the negative sentiment toward ORG. This may be because in many cases within the dataset where the model was trained, the entity of interest and the sentiment toward it are introduced at the beginning of the article. Text-davinci-003, however, successfully captured the sentiment, likely due to its ability to better understand context over longer stretches of text. Overall, XLM-RoBERTa emerges as the top performer, due to its fine-tuning on the specific distribution of the dataset. Text-davinci-003 tends to overpredict minority classes but appears to excel in managing more complex examples.

In terms of the performance of the evaluated models across the three languages of the dataset, Italian consistently shows better results compared to German and Ladin (Table 11). However, when focusing on the F_1-macro-pos-neut score calculated on the positive and neutral categories, we see that for these two most represented categories in the dataset, the measured accuracy values are in line between the three languages. The absence of negative training examples for Ladin, while there are a handful in the test set, explains the lower macro-averaged F1 score across all three categories for Ladin.

The hypothesis that sentiment analysis of news articles is more challenging than sentiment analysis of customer reviews seems to be confirmed by our experiments. XLM-RoBERTa achieved significantly lower results (F_1-macro: 58.32, F_1-macro-pos-neut: 77.28) than when the same model was evaluated on the customer reviews (F_1-macro: 64.00, F_1-macro-pos-neut: 95.50).

Our findings suggest that the assessed dataset can be used to train models for three main purposes. Firstly, the models can be employed to automate the annotation of news articles for many organizations monitored by the Company without requiring further manual review. This is feasible due to the relatively low misclassification rate of positive examples as negative (and vice versa) by the models, as shown in Table 10. For many clients, this level of error might be acceptable, allowing sentiment annotation to be automated for those cases. Secondly, since the accuracy of classifying articles as Neutral is approximately 90% for entities in the dataset, it is reasonable to trust the classifier prediction when it identifies an article as Neutral, which happens in 78% of cases. In cases where the classification differs, it is advisable to verify the automatic annotation by expert annotators. This implies that only 22% of documents (from 44 articles to 66 articles per day) need verification, resulting in a time-saving of 78% for annotation. Thirdly, these models can accelerate the manual verification process, focusing solely on articles associated with entities where the model accuracy on the Neutral category falls below 85%, which applies to a minority of entities (Fig. 1). In this scenario, the time saved for annotation is at least 50%, reducing the current 400 to 600 minutes, to 200 to 300 minutes.

Although our findings suggest that the machine learning models evaluated in this work could improve annotation efficiency, we plan to conduct a human study to confirm whether these models can indeed accelerate the annotation process in practice.

9. Conclusion

The study investigated the effectiveness of machine learning in sentiment analysis of multilingual news articles in languages with fewer resources than English. Our findings confirm previous studies, indicating that sentiment analysis of news articles is more difficult than that of other domains such as customer reviews, owing to the complex language and indirect expressions of sentiment. Although the dataset used in the study cannot be shared publicly for privacy reasons, the results provide valuable insights into the challenges and opportunities of sentiment analysis in news articles.

The use of machine learning has shown the potential to reduce the manual annotation process, which may involve automating annotation for either all news articles or a significant portion of them.

To address the imbalance in the current dataset, future research will investigate the feasibility of using negative articles from entities like those already included in the dataset. This approach could potentially enhance the identification of negative sentiment associated with entities within the dataset, even in the absence of directly related negative articles. Furthermore, there is currently ongoing research on leveraging large language models like Modello Italia for sentiment analysis. Additionally, evaluating Llama-3-8B with few-shot prompting holds promise. This technique requires careful selection of representative examples but could potentially address the model's output inconsistency issues.

10. Acknowledgments

This work was partially supported by the RecoFeel Project funded by Provincia Autonoma di Bolzano, Ufficio Innovazione, Ricerca e Università under the law n. 14/2006 (Act n. 5/2021).

11. References

References

Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3), June.

Al Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine El Kadiri. 2018. Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127(C):511–520, May.

Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Barbieri, Francesco, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy, December.

Barnes, Jeremy, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States, July. Association for Computational Linguistics.

Basile, Pierpaolo, Valerio Basile, Danilo Croce, and Marco Polignano. 2018. Overview of the EVALITA 2018 aspect-based sentiment analysis task (ABSITA). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Basile, Valerio, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy, December.

Bastan, Mohaddeseh, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's sentiment prediction. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 05.

Bonadiman, Daniele, Giuseppe Castellucci, Andrea Favalli, Raniero Romagnoli, and Alessandro Moschitti. 2017. Neural sentiment analysis for a real-world application. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, December.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Cui, Jingfeng, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. 2023. Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, 56(8):8469–8510, January.

Dang, Nhan Cach, María N. Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3).

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*, page 519–528, Budapest, Hungary, May. Association for Computing Machinery.

De Mattei, Lorenzo, Graziella De Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, and Giulia Rambelli. 2020. Ate_absita @ EVALITA2020: overview of the aspect term extraction and aspect-based sentiment analysis task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Du, Chunning, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online, July. Association for Computational Linguistics.

Fang, Xing and Justin Zhijun Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data*, 2:1–14.

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

Godbole, Namrata, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 219–222, Boulder, Colorado, USA, March. Association for the Advancement of Artificial Intelligence (AAAI).

Hamborg, Felix and Karsten Donnay. 2021. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online, April. Association for Computational Linguistics.

Huq, M. Rezwanul, Ahmad Ali, and Anika Rahman. 2017. Sentiment analysis on twitter data using knn and svm. *International Journal of Advanced Computer Science and Applications*, 8.

Jurek, Anna, Maurice D. Mulvenna, and Yaxin Bi. 2015. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4:1–13.

Kang, Hanhoon, Seong Joon Yoo, and Dongil Han. 2012. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010.

Li, Yung-Ming and Tsung-Ying Li. 2013. Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1):206–217.

Li, Zheng, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 2237–2243, Melbourne, Australia, August. International Joint Conferences on Artificial Intelligence Organization.

Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge, United Kingdom.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

McDonald, Ryan, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 432–439, Prague, Czech Republic, June. Association for Computational Linguistics.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date. [online] Available at: `https://ai.meta.com/blog/meta-llama-3/` [Accessed December 2024].

Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June. Association for Computational Linguistics.

Pak, Alexander and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Pang, Bob and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pontiki, Maria, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Rana, Shweta and Archana Singh. 2016. Comparative analysis of sentiment orientation using svm and naive bayes techniques. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 106–111, Dehradun, India, October. IEEE.

Rijsbergen, Cornelis Joost Van. 1979. *Information Retrieval*. Butterworth-Heinemann, 2nd edition.

Rosenthal, Sara, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August. Association for Computational Linguistics.

Singh, Vivek Kumar, Rajesh Piryani, Ashraf Uddin, and Pranav Waila. 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717, Noida, India. IEEE.

Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Song, Youwei, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Targeted sentiment classification with attentional encoder network. In Igor V. Tetko, Věra Kůrková, Pavel Karpov, and Fabian Theis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, pages 93–103, Munich, Germany, September. Springer International Publishing.

Sprugnoli, Rachele, Francesco Mambrini, Marco Carlo Passarotti, and Giovanni Moretti. 2023. The sentiment of Latin poetry. Annotation and automatic analysis of the odes of Horace. *Italian Journal of Computational Linguistics*, 9:53–71.

Sun, Chi, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tan, Kian, Chin-Poo Lee, and Kian Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13:4550, 04.

Turney, Peter D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 417–424, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Wiebe, Janyce M., Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA, June. Association for Computational Linguistics.

Wnkhade, Mayur, Annavarapu Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55:1–50, 02.

Yadav, Ashima and Dinesh Kumar Vishwakarma. 2019. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53:4335 – 4385.

Ye, Junjie, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *ArXiv*, abs/2303.10420.

Zeng, Biqing, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. LCF: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9:3389, 08.

Zhang, Bowen, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online, July. Association for Computational Linguistics.

Zhou, Xinjie, Xiaojun Wan, and Jianguo Xiao. 2016. Attention-based LSTM network for cross-lingual sentiment classification. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 247–256, Austin, Texas, November. Association for Computational Linguistics.