data

Review

The Modern Greek Language on the Social Web: A Survey of Data Sets and Mining Applications

Maria Nefeli Nikiforos *,†, Yorghos Voutos *,†, Anthi Drougani, Phivos Mylonas and Katia Lida Kermanidis

Department of Informatics, Ionian University, 49132 Corfu, Greece; p14drou@ionio.gr (A.D.); fmylonas@ionio.gr (P.M.); kerman@ionio.gr (K.L.K.)
* Correspondence: c19niki@ionio.gr (M.N.N.); george.voutos@gmail.com (Y.V.)
† These authors contributed equally to this work.

Abstract: Mining social web text has been at the heart of the Natural Language Processing and Data Mining research community in the last 15 years. Though most of the reported work is on widely spoken languages, such as English, the significance of approaches that deal with less commonly spoken languages, such as Greek, is evident for reasons of preserving and documenting minority languages, cultural and ethnic diversity, and identifying intercultural similarities and differences. The present work aims at identifying, documenting and comparing social text data sets, as well as mining techniques and applications on social web text that target Modern Greek, focusing on the arising challenges and the potential for future research in the specific less widely spoken language.

Keywords: social web language; modern greek; natural language processing; data mining; machine learning; text analysis

check for updates

Citation: Nikiforos, M.N.; Voutos, Y.; Drougani, A.; Mylonas, P.; Kermanidis, K.L. The Modern Greek Language on the Social Web: A Survey of Data Sets and Mining Applications. *Data* **2021**, *6*, 52. https://doi.org/10.3390/data6050052

Academic Editor: Erik Cambria

Received: 28 February 2021
Accepted: 10 May 2021
Published: 17 May 2021

1. Introduction

Over recent years, social web text (also known as *social text*) processing and mining has attracted the focus of the Natural Language Processing (NLP), Machine Learning (ML) and Data Mining research communities. The increasing number of users connecting through social networks and web platforms, such as Facebook and Twitter, as well as numerous Blogs and Wikis, creates continuously a significant volume in written communication through the Web [1–7]. The amount and quality of information and knowledge extracted from social text has been considered crucial to studying and analyzing public opinion [1,3,5,8,9], as well as linguistic [2,7,10–15] and behavioral [4,6,16–18] patterns. In its typical form, social text is often short in length, low in readability scores, informal, syntactically unstructured, characterized by great morphological diversity and features of oral speech, misspellings and slang vocabulary, consequently presenting major challenges for NLP and Data Mining tasks [2,4,7,10,11,13,13–16,19]. Therefore, several works have attempted to develop tools to extract meaningful information from this type of text with applications in numerous fields, such as offensive behavior detection, opinion-mining, politics analysis, marketing and business intelligence, etc. Capturing public sentiment on matters related to social events, political movements, marketing campaigns, and product preferences passes through emotion processing methodologies, which are being developed in the inter-compatible Web. On that notion, the combination of several academic principles (inter-disciplinarity), allows experts to develop "affect-sensitive" systems through syntax-oriented techniques (e.g., NLP) [20].

ML tools and techniques have been significant in NLP and Data Mining tasks on social text, due to their adaptability to the data, as well as their ability to efficiently handle vast volumes of data. "ML is programming computers to optimize a performance criterion using example data or past experience" [21]. During the learning phase, parameters of a general model are adjusted according to the training data. During the testing phase, the specialized model is tested with new, not previously known data, and its performance regarding a target task is evaluated [21,22]. The objective of supervised learning is to map the provided input to an output, where true values are acquired by a supervisor [21,22]. The objective of unsupervised learning is to detect the regularities in the provided input and its underlying structure, though the true values of the output are not acquired by a supervisor [21,22]. Semi-supervised learning includes training with both labeled and unlabeled data [21]. In reinforcement learning, an agent learns behavior through trial-and-error in a dynamic environment [23]. It is applied when the target task results from a sequence of actions [21,22,24,25].

There are several evaluation metrics to evaluate the performance of ML algorithms [13]. At this point we focus on popular metrics related to text mining as they have been used in the literature [2,4,6,15,26] of this survey. Accuracy is calculated to show the correctness of the prediction for the data examples of all classes. Precision describes the ratio of total classifications made by the algorithm on the test set that are correct, while recall is the ratio of the total test data that are correctly identified by the classifier [27]. The F1 score, also known as F1-measure[1], occurs by calculating the harmonic mean of precision[2] and recall[3].

More specifically, recall is referred to the sensitivity rate and precision is referred to the Positive Predictive Value (PPV) of each measure. Furthermore, the F1-score employs the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) metrics to measure the test's expectation (positive or negative) and subjective observation (true or false). The related metrics presented in the equations above, compare the classifier including true negative rate and accuracy [27]. True negative rate is also called specificity and is expressed by True Negative Rate (TNR) equation[4].

Though most of the reported work is on social text written in widely spoken languages, such as English, the significance of approaches that deal with less commonly spoken languages is evident for reasons of preserving and documenting minority languages, cultural and ethnic diversity, identifying intercultural similarities and differences. Modern Greek is a morphologically rich language, which has a variety of potential grammatical forms for each word root, i.e., words may have different inflected forms, due to different grammatical properties [2,4,12,28]. Another usual phenomenon is that a root with different suffixes could create morphologically similar words, though with significantly different meaning, resulting in numerous, semantically different, derivative words [2,4,5]. Although the Greek language is a branch of the Indo-European family [29], there are differences between the related languages [30], especially in their morphological characteristics. Therefore, linguistic idiom is a feature of Greek social text that needs to be studied separately. Nevertheless, researchers have studied the reaction patterns of social media users in the language of their interest, such as Czech [31], Arabic [32], as well as in different languages [33] on various applications, including sentiment analysis, regarding either a single language or a multilingual setting [19,20,34].

The Modern Greek alphabet consists of symbols that differ from those of languages that derive from the Latin, and thus the existing NLP tools for those languages are often not easily adapted and applied to Modern Greek text [35]. Additionally, the existence of dialects and idioms, such as Grico, Cretan, Cypriot and Pontic Greek, and their common use, add to the complexity of the language [2,7,36]. Clitics in Modern Greek may also have varying morphological and syntactic properties, depending on the dialect [36]. These characteristics, along with the existence of limited linguistic processing resources, data sets and NLP tools, have rendered NLP, ML and Data Mining on text written in the Modern Greek language to be quite challenging in previous work [2,4–7,12,35]. The present work aims at identifying, documenting and examining social text data sets in Modern Greek.

[1] $F1 = \frac{TP}{TP+\frac{FP+FN}{2}}$.

[2] $Recall = \frac{TP}{TP+FN}$.

[3] $Recall = \frac{TP}{TP+FN}$.

[4] $TNR = \frac{TN}{TP+FN}$.

State-of-the-art mining techniques and applications on social text that target Modern Greek are also identified and examined in this work, focusing on the arising challenges and the potential for future research in the specific less widely spoken language.

The structure of this paper is the following: Section 2 discusses the recent literature on linguistic and behavioral patterns, regarding both linguistic patterns analysis (Section 2.1) and offensive behavior and language detection (Section 2.2). Section 3 analyzes the recent literature on opinion-mining, regarding both politics and voting analysis (Section 3.1) and marketing and business analysis (Section 3.2). Sections 4 and 5 discuss the findings, conclude the paper and outline guidelines for future work. All abbreviations are presented at the end of the document.

2. Linguistic and Behavioral Patterns

The identification of patterns in data has been a demanding task in the context of social text, mainly due to its unstructured nature, rich morphology and increasing volume [2,4,7,10,11,13–16]. Several researchers have focused on the identification of linguistic and/or behavioral patterns of interest in social text data. The most commonly used process is the following: At first, the data are collected from the social web, usually by a web scraper or through an Application Programming Interface (API). Then, they are preprocessed, including normalization and transformation, and encoded into a data set with a form and structure suitable for the stage of processing; the implementation of Data Mining and NLP techniques. At the next stage, experiments with the data set and several ML algorithms are conducted. Finally, the results are interpreted and the performance of the algorithms is evaluated.

In this Section, the recent literature on linguistic and behavioral patterns is discussed, regarding both linguistic patterns analysis (Section 2.1) and offensive behavior and language detection (Section 2.2).

2.1. Linguistic Patterns Analysis

There are several approaches that have attempted to identify, analyze and extract linguistic patterns by developing and using various NLP tools [2,12]. Other work focuses on the creation of corpora from various linguistic contexts to apply either classification [7], or machine translation [37]. Additionally, there are certain approaches that have explored argument extraction and detection from text corpora [13–15]. Another approach attempted authorship attribution and author's gender identification for bloggers [10,11]. An overview of the recent literature regarding linguistic patterns analysis, which is discussed in this subsection, is shown in Tables 1 and 2.

Table 1. Overview of the literature (linguistic patterns analysis). Social media, data sets and corpora, methods applied on data, and the resulting tool.

Paper	Social Media	Data	Methods	Tool
[2]	Twitter	2405 tweets 31,697 tokens (April 2019)	tokenization, normalization, encoding, annotation	POS tagger
[12]	Twitter	4,373,197 tweets 30,778 users 54,354 hashtags (April 2008–November 2014)	automated & manual rating, removal: stop words & tone marks, stemming, uppercase	Sentiment analysis lexicon
[7]	Twitter Facebook forums, blogs	1039 sentences 7026 words (Cypriot Greek) 7100 words (Modern Greek) (March–April 2018)	anonymization, manual annotation, removal: tabs, newlines, duplicate punctuation, insertion: spaces, n-grams, encoding, tokenization	Bidialectal classifier
[37]	MOOC	multilingual corpus course forum text quiz assessment text subtitles of online video lectures	conversion into plain text, removal: special characters, non-content lines, multiple whitespaces, tokenization, sentence segmentation, special elements markup	-
[13]	Twitter, news blogs, sites	204 documents 16,000 sentences: 760 argumentative	manual annotation, tokenization, sentence splitting, POS tagging, feature selection, gazetteer lists, lexica, TF-IDF	-
[14]	Twitter, news blogs, sites	204 documents 16,000 sentences: 760 argumentative comparison with NOMAD data set	manual annotation, tokenization, sentence splitting, POS tagging, feature selection, gazetteer lists, lexica, TF-IDF	-
[15]	Twitter Facebook news, blogs	1st: 77 million documents 2nd: 300 news articles, 1191 argumentative segments	POS tagging, cue words, distributed representations of words, feature extraction, sentiment analysis, lowercase	-
[10]	Blogs	1000 blog posts 406,460 words (September 2010–August 2011)	stylometric variables, character & word uni-grams, bi-grams, tri-grams, feature extraction	Authorship attribution & author's gender identification
[11]	Twitter	45,848 tweets	removal: stop words, encoding: Bag-of-Words, TF-IDF	Author's gender identification

Table 2. Overview of the literature (linguistic patterns analysis). Machine learning and other algorithms, experimental results, contribution, and open issues.

Paper	Algorithms	Results	Contribution	Open Issues
[2]	Naive Bayes ID3	accuracy up to 99.87%	1st data set for Greek social text 1st tag set 1st supervised POS tagger	larger data sets data from different social media syntactic & semantic analysis tools linguistic diversity by region tracking controversial events & mapping connections with users
[12]	Pearson Kendall correlation	sentiment correlation	public benchmark data set set of intensity rated tweets automated method for detecting intensity (tweets & hashtags) temporal changes in intensity (hashtags)	lexicon for social text more linguistic data larger data set & number of raters
[7]	Naive Bayes, SVM, LR	95% mean accuracy	1st classifying Greek dialects in social text bidialectal corpus & classifier most informative features	applications in social media moderation and academic research larger corpus including POS detecting dialects prior to online translation extension with Greeklish, Pontic & Cretan Greek distinction between Katharevousa & Ancient Greek
[37]	-	-	multilingual parallel corpus to train, tune, test machine translation engines translation crowd-sourcing experiment examination of difficulties: text genre, language pairs, large data volume, quality assurance, crowd-sourcing workflow	-
[13]	LR, RF, SVM, CRF	accuracy up to 77.4%	2-step argument extraction novel corpus most determinant features	more features & algorithms testing of Markov models

Table 2. *Cont.*

Paper	Algorithms	Results	Contribution	Open Issues
[14]	LR, RF, SVM, CRF	accuracy up to 77.4%	2-step argument extraction novel corpus most determinant features	more features & algorithms testing of Markov models comparing performance with approaches for English experiments with unsampled data
[15]	word2vec CRF	up to 39.7% precision 27.59% recall 32.53% F1 score	semi-supervised multi-domain method argument extraction novel corpus	extending the gazetteer lists bootstrapping on CRF more algorithms patterns based on verbs and POS grammatical inference algorithm
[10]	SVM	accuracy 85.4% & 82.6%	tool for authorship attribution & author's gender identification with many candidates novel social text corpus 10 most determinant features	-
[11]	SVM	accuracy up to 70%	novel, manually annotated, corpus NLP framework for gender identification of the author	more featurescombining gender & age neural networks

2.1.1. NLP Tools

In a previous approach, Nikiforos & Kermanidis [2] created the first annotated data set for social text in Modern Greek (publicly available[5]), the first tag set for the annotation of this data set, regarding the Part-of-speech (POS) in Modern Greek, and developed the first supervised POS tagger for social text in Modern Greek, which achieved considerably high performance. The novel data set consisted of 2405 tweets from April 2019, tokenized in 31,697 separate tokens. Certain types of tokens (hyperlinks, punctuation marks, symbols, etc.) were replaced with code-words to be normalized. The data set was annotated manually according to the proposed tag set, which consisted of 22 distinct tags, including special tags for Twitter language specifics (hashtags, at-mentions, emoticons, hyperlinks).

Modern Greek grammars and a set of guidelines were provided to the annotators to aid them in their task, especially in ambiguous cases. During annotation, it was observed that traits of oral speech, gluing of two or more words, idioms and dialects were more common in tweets from unofficial accounts[6] than in those from public figures' accounts. The researchers provided a detailed description and disambiguation of the types of tokens and tags, and their respective frequencies. For the ML experiments, a learning example was created for each token, with its 3-letter suffix and its neighboring words (3 preceding and 3 following) and their respective tags as features, and its tag as the class, resulting in 31,697 learning examples, 13 features and 22 classes.

Experiments were conducted with 80% of the data set considered to be the training set and the remaining 20% considered to be the test set, using the RapidMiner Studio[7]. Naive Bayes classifier achieved 99.87% accuracy, precision from 96.6% to 100%, recall from 99.41% to 100% and F1 score from 98.27% to 100%. Punctuation marks, expressions, symbols, emoticons and interjections were classified correctly more easily compared to the other tags. The ID3 classifier achieved 99.44% accuracy, precision from 98.03% to 100%, recall from 96.09% to 100% and F1 score from 97.61% to 100%. Punctuation marks, expressions, symbols and emoticons were classified correctly more easily compared to the other tags. Hashtags were classified correctly less easily by both classifiers, as they may often be used in the place of any token.

Additionally, the wrong predictions of both classifiers were identified. In particular, both classifiers confused: articles and pronouns with similar suffixes, at-mentions and hashtags at the end of the tweet, conjunctions and adverbs or particles at the beginning of the tweet, adverbs and particles or pronouns or prepositions or verbs with similar suffixes, pronouns and particles or articles with similar suffixes, or prepositions or common nouns at the beginning of the tweet. This work, although rather pioneer in its nature, allowed also for several future optimizations, namely conducting experiments with a larger data set to address the issue of overfitting, creating data sets of social text in Modern Greek from different social networks, developing tools for syntactic and semantic analysis for this type of data, further exploring of the linguistic diversity of tweets according to geographic region, and tracking controversial events and mapping their connections with user accounts.

Kalamatianos et al. [12] created a publicly available[8] benchmark data set of tweets written in Modern Greek, a set of manually rated tweets regarding sentiment intensity, and presented a method for automatic detection of the sentiment intensity of tweets, as well as of the sentiment rating of hashtags. They also examined the variations of the intensity of "Happiness" and "Anger" per hashtag through time in correlation with actual events. The novel data set consisted of 4,373,197 tweets from 30,778 users of the same user network (following-follower relationship), containing 54,354 distinct hashtags, which were collected from April 2008 to November 2014.

[5] https://hilab.di.ionio.gr/index.php/en/datasets/, (accessed on 15 November 2020).

[6] Refers to accounts that are not directly related to specific individuals, usually public figures.

[7] https://rapidminer.com/, (accessed on 18 February 2021).

[8] http://hashtag.nonrelevant.net/downloads.html, (accessed on 18 February 2021).

In this framework, tweets were categorized according to the 41 most commonly used hashtags. During preprocessing, stop words and tone marks were removed, all letters were converted to uppercase and a stemmer was applied. Eventually, 6 sentiments were considered, namely: "Anger", "Disgust", "Fear", "Happiness", "Sadness" and "Surprise". The raters rated manually a subset of the data set, consisting of 681 tweets filtered by 10 hashtags, in a scale of 0 to 5 per sentiment. The inter-rater agreement was calculated by Pearson's linear correlation coefficient; it was moderate for "Fear" (0.415), "Happiness" (0.477), "Sadness" (0.530) and "Surprise" (0.398), while ratings for "Anger" and "Disgust" were not correlated (0.064 and -0.034, respectively), due to the ambiguity caused by the use of sarcasm.

The Greek Sentiment Lexicon, consisting of 2315 entries along with metadata about the tone, the POS, the objectivity and the emotional content of each word, was used for the automated rating. Both the inter-rater agreement and the pairwise correlation of the sentiments were calculated by Pearson's linear correlation coefficient; "Anger" was highly correlated with "Disgust" (0.827) and "Happiness" was highly correlated with "Surprise" (0.558). However, merely 11.7% of the words of tweets existed in the lexicon. The tweet sentiment rating was calculated with 4 formulae, which combined tweet and lexicon information. The hashtag sentiment rating was calculated with 2 formulae, which combined tweet and hashtag information. Both types of rating were evaluated by comparing the automated rating to the raters' rating with the Pearson and the Kendall correlation; "Sadness" and "Surprise" were not correlated, the correlation was fair for "Happiness" and moderate for "Fear". The researchers examined the variations of the intensity of "Happiness" and "Anger" through time for specific hashtags, by calculating daily the quadratic mean of sentiment words per tweet. The proposed approach leaves room for optimization among the tasks of developing a lexicon specialized for social text, considering more linguistic data (e.g., POS, emoticons, punctuation), extending the benchmark data set and increasing the number of raters to improve the results for "Anger" and "Disgust", as well as further evaluating sentiment variations through time.

2.1.2. Linguistic Classification and Corpora

In the field of linguistic classification, Sababa & Stassopoulou [7] were the first to attempt Greek dialect classification, by distinguishing Cypriot Greek from Modern Greek social text. They created a bidialectal corpus of social text, developed a classifier, by extracting n-gram features and testing multinomial Naive Bayes, Logistic Regression and Support Vector Machine (SVM), and discovered the most informative features. The corpus and the source code for collecting the data and building and testing of the classifiers, are publicly available[9]. The corpus data consisted of 1039 sentences, manually annotated, and collected by filtering posts and comments from Twitter, Facebook, forums and blogs from March to April 2018. Cypriot Greek data and Modern Greek data contained 7026 words and 7100 words, respectively.

All data were anonymized, while Greek text written in Latin letters (Greeklish), or containing any numbers or Latin letters, or using Cypriot diacritics was not considered. During preprocessing, consecutive whitespaces were replaced by a single space, to remove tabs and newlines, duplicate consecutive punctuation marks were removed, and spaces were inserted where necessary. The sentences were extracted with NLTK's sentence tokenizer, while hyperlinks, Hypertext Markup Language—HTML entity references, Twitter-specific tags and keywords, punctuation and Unicode characters were removed, and all letters were converted to lowercase. The study of the most commonly used words, n-grams and characters showed notable differences between the two dialects; the 10 most frequent uni-grams and bi-grams were unique to Cypriot Greek, and the most frequent characters, as well as word order and syntax, were different.

[9] https://github.com/hb20007/greek-dialect-classifier, (accessed on 15 November 2020).

During feature extraction, word uni-grams and bi-grams, and padded character uni-grams, bi-grams and tri-grams (to indicate prefix or suffix of the word) were extracted from the sentences with nltk.ngrams, and encoded in feature vectors with Scikit-learn[10] to address overfitting. A total of 80% of the data was considered to be the training set and the remaining 20% was considered to be the test set. Multinomial Naive Bayes, SVM and Logistic Regression were built, trained and tested with Scikit-learn. The training-testing process for each classifier was repeated for 5 times with different partitions and the mean accuracy was calculated. The best results were obtained with multinomial Naive Bayes (95% mean accuracy, 96% F1 score), compared to SVM and Logistic Regression (92% and 94% mean accuracy, respectively). Experiments with different feature subsets were also conducted, the main results being: character n-gram features with padding were significant for the successful classification, word uni-gram and bi-gram features improved accuracy, while ignoring infrequent words did not improve accuracy.

Additionally, the researchers discovered which features were more informative (highest weights) for multinomial Naive Bayes and thus determinant for the classification; Cypriot Greek features had higher weight. Error analysis highlighted the need for a larger data set with POS features. All in all, this research work may be further extended by applying the proposed classifier in the framework of social media moderation and academic research, using the classifier to identify the dialect prior to online translation, by creating a larger data set with POS features, and by developing an extension for including Greeklish, Pontic and Cretan Greek, by distinguishing between Katharevousa and Ancient Greek.

Under a different perspective, Sosoni et al. [37] created a multilingual parallel corpus to train, tune and test machine translation engines, and conducted a translation crowdsourcing experiment. They examined the difficulties regarding the text type, the vast volume of data, the number of language pairs, the quality assurance, and the crowdsourcing workflow issues. The corpus consisted of informal text (course forum text) and formal text (quiz assessment text, subtitles of online video lectures) written in 11 distinct languages, including Modern Greek, from Massive Open Online—MOOC type courses (Iversity.org, Videolectures.NET, Coursera, QED) of various topics.

During the preparation stage, all data were converted into plain text (UTF-8[11] encoded) with UNIX-based shell and Python scripts, and special characters, non-content lines, and multiple whitespaces were removed. Then, the data were tokenized and segmented[12] to words or punctuation, and sentences, respectively. Certain special elements that were not considered for translation (e.g., URLs, emoticons) were marked up with tags. The CrowdFlower[13] platform was used to translate both formal and informal educational content segments in English to both low- and high-resource languages, in combination with quality measures and features. A total of 2050 workers contributed to the translation (March to June 2017), according to given guidelines regarding linguistic and formatting ambiguities. The accuracy of each worker was evaluated, to ensure high quality translations, and accurate workers were labeled as trusted. The formal text consisted mainly of domain specific terms and scientific formulae, making the translation process quite challenging for the workers, and subtitles, which contained spontaneous speech features (e.g., unfinished sentences, elliptical formations, repetitions, interjections). The tuning and testing set consisted of 5000 segments, translated by 2 to 3 workers per target language. Regarding Modern Greek, over 70,000 segments were translated, a significant amount, considering it is a low-resource language. Less than 5000 judgements were not trusted, while 90% of the workers originated from Greece.

[10] https://sklearn.org/, (accessed on 15 November 2020).

[11] Unicode Transformation Format.

[12] https://github.com/mixstef/tramooc, (accessed on 19 November 2020).

[13] https://appen.com/, (accessed on 19 November 2020).

2.1.3. Argument Extraction

Goudas et al. [13] proposed a two-step methodology for argument extraction from social text written in Modern Greek. The corpus consisted of 204 documents (16,000 sentences: 760 argumentative), manually annotated and written in Modern Greek, collected from Twitter, news, blogs and sites, and concerned renewable energy sources. Domain entities represent the claims, and educational content segments in favor or against them. The Ellogon [26] language engineering platform was used for tokenization, sentence splitting, POS tagging, as well as for the generation of feature vectors. The first step of the proposed methodology was the classification of the sentences into those which contained arguments and those which did not, with supervised learning algorithms in Waikato Environment for Knowledge Analysis (WEKA)[14], such as Logistic Regression, RF, SVM and Naive Bayes. The goal of this step was to identify which features of the sentences, from both the state-of-the-art literature (10 features: position, comma token number, connective number, verb number, number of verbs in passive voice, cue words, domain entities number, adverb number, word number, and word mean length) and the novel features of social text (5 features: adjective number, entities in previous sentences, cumulative number of entities in previous sentences, ratio of distributions, and distributions over uni-grams, bi-grams, tri-grams of POS tags), achieved the best classification results.

The second step was the identification of the exact fragments of sentences that contained arguments with Conditional Random Fields (CRF) in Ellogon. The feature set consisted of the words in the argumentative sentences, gazetteer lists of entities according to topic, as well as of indicator sentences and cue words, and verb and adjective lexica by Term Frequency-Inverse Document Frequency (TF-IDF) between the argumentative and the non-argumentative sentences. Baseline classifiers with 10-fold cross-validation showed poor results for both steps. For the first step, Logistic Regression, RF, SVM and Naive Bayes were evaluated both with 10-fold cross-validation on a sampled, balanced data set (Logistic Regression: up to 77.4% accuracy, 77.4% precision and 77.3% recall), and with splitting the initial data set in 70% training set and in 30% unsampled test set (49% accuracy). For the second step, the data set consisted of sentence segments that contained arguments and the CRF achieved 62.23% precision, 32.43% recall and 42.37% F1 score with 10-fold cross validation.

In a more recent work of the same authors [14], they applied their two-step methodology for argument extraction from social text written in Modern Greek as a module to the NOMAD[15], a platform for policy making. They compared the results of their previous work with those of experiments conducted with a data set extracted from NOMAD. The researchers made several suggestions for future work: (a) examining more features, such as verbal tense and mood, for the first step to improve accuracy, (b) experimenting with more sophisticated ML algorithms for argument identification, (c) testing Markov models and additional features for the argumentative segment extraction, (d) comparing performance with the state-of-the-art approaches for English, and (e) conducting experiments with unsampled data.

Sardianos et al. [15] attempted the identification of Modern Greek social text segments that contained arguments, by creating a novel, manually annotated corpus[16] and applying semi-supervised learning with CRF. The proposed approach focused mostly on e-Government and the policy making domain. The data consisted of text written in Modern Greek, and originated from Twitter, Facebook, news and blogs, including various topics (e.g., politics, economics, culture, sports). Several features were extracted (words, POS tags, cue words, distributed representations of words, sentiment analysis). The first set of experiments were conducted with a word2vec model[17] to identify similar words and a

[14] http://www.cs.waikato.ac.nz/~ml/weka, (accessed on 23 December 2020).

[15] http://nomad-project.eu/en, (accessed on 23 December 2020).

[16] Only the corpus containing news can be redistributed for research purposes.

[17] A technique for NLP that employs a two-layer neural net that processes text by creating a vector of real numbers to represent a word.

corpus, consisting of 77 million documents. All letters were converted to lowercase. The obtained results highlighted that data from news and blogs produced more fine-grained and efficient models, compared to those produced by data from Facebook and Twitter, which are more unstructured and noisier, and their vocabulary was not always relevant to a specific topic.

In this work, CRF with 10-fold cross-validation is implemented as a basic validation mode, thus setting a benchmark for model's performance and for selecting the characteristics (features) that enhance its performance. Knowing the context can provide higher accuracy. During the first phase achieving the optimal performance is not required, but rather providing a baseline performance of a simple model. Sardianos et al. used the Beginning—Inside—Outside—Unit (BILOU) representation; there are 5 different tags, corresponding to 5 different classes. More specifically, when 2 classes are used for classification, the metrics (Precision, Recall, F1) must be over 50% to outperform randomness. However, when 5 classes are used for classification, the threshold for randomness is set at 20%.

For the evaluation of the first set of experiments, a set of 20 words (not topic-specific) and a list of the 5 most similar words for each word, obtained from the word2vec model, were given to 2 human annotators. The inter-annotator agreement was 0.825 for entities and 0.850 for cue words, concluding that word2vec could contribute to the expansion of the cue word lexica. The second set of experiments was conducted with CRF and distributed representations of words and a novel, manually annotated corpus, consisting of annotated segments that include arguments from 300 news articles. Two annotators with specific instructions identified positive and negative claims from 150 articles each. When 150 articles in total were annotated (first stage of annotation), pre-annotation was applied, with CRF detecting 4524 segments. At the second stage of annotation, the annotators had to evaluate the segments detected by the CRF as argumentative, as well as to revise their annotations and correct any errors. The final corpus consisted of 1191 argumentative segments.

For the evaluation of the second set of experiments, the CRF model was tested with 10-fold cross-validation and words and POS tags as features. Each token of each sentence was classified either as the beginning of or internal to or the ending of the argument, or as external of the argument. The best results were obtained with 2 or 5 words used as context, achieving up to 39.7% precision, 27.59% recall and 32.53% F1 score. The researchers made several suggestions for future work: (a) extending the gazetteer list of entities and cue words, (b) applying bootstrapping techniques on CRF, (c) experimenting with different classification algorithms, and d. extracting patterns based on verbs and POS and applying a grammatical inference algorithm.

2.1.4. Authorship Attribution and Gender Identification

Mikros [10] created a tool for authorship attribution and author's gender identification with many candidates, with a novel social text corpus written in Modern Greek from September 2010 to August 2011. They also identified the 10 most determinant features for the author's gender identification, as well as for the authorship attribution. The corpus consisted of 1000 blog posts (50 blog posts per author) or 406,460 words, of the topic personal affairs; posts of 10 male and 10 female authors were considered. Several standard stylometric variables, regarding vocabulary richness, word length, and letter frequencies, and the 300 most frequent character and word uni-grams, bi-grams and tri-grams were calculated. As a result, the final feature set consisted of 1356 features. SVM with 10-fold cross-validation achieved 82.6% accuracy for the author's gender identification, and 85.4% accuracy for the authorship attribution. They concluded that the performance is excellent for up to 4 candidates, author gender was identified by syntactic and morphological patterns, and authorship was connected to specific high-frequency words.

Baxevanakis et al. [11] created a novel, manually annotated, social text corpus[18] written in Modern Greek and presented an NLP framework for gender identification of the

[18] Corpus and code can be provided upon request.

author. The data originated from a random population of Twitter users; 463 authors with 99.023 mean number of tweets per author, resulting in 45,848 tweets. Regarding annotation, in the cases of annotator disagreement about the gender of the author, the specific author and their tweets were removed from the data set. During the preprocessing stage, the stop words were removed. Bag-of-Words[19] was used for the encoding. The feature set consisted of 6 features, containing information regarding Twitter handle, name and gender of the author, Twitter issued profile identification number, and tweet link and text. The experiments were conducted with both balanced and imbalanced data, using the scikit-learn python library, and included multinomial Naive Bayes, SVM, k-Nearest Neighbor (kNN) and Random Forest (RF). SVM and TF-IDF showed the best results, achieving up to 70% accuracy. The researchers made several suggestions for future work: (a) using more features, e.g., image data, statistics of tweeting behavior, (b) combining gender and age, and (c) experimenting with neural networks.

2.2. *Offensive Behavior and Language Detection*

There are several approaches that have attempted to detect and analyze bullying and aggressive behavior in Virtual Learning Communities (VLCs) [4,16,17]. Other work focuses on offensive language identification and analysis in tweets [6,18]. An overview of the recent literature regarding offensive behavior and language detection, which is discussed in this subsection, is shown in Tables 3 and 4.

2.2.1. Bullying in VLCs

Nikiforos et al. [4] were the first to study the influence of VLCs on behavior modification regarding bullying, with a NLP and ML framework for automatic detection of aggressive behavior and bullying in Modern Greek text and authentic humanistic data collected under real conditions, addressed to teachers for aiding them to intervene when necessary. More specifically, they examined if the behavior of individuals that bullied in their Physical Learning Community (PLC) could be modified, when integrated in a VLC. The PLCs were virtualized with the Wikispaces web-based collaborative learning environment, and artifacts and dialogues were collected from its log files. Two VLCs were examined; VLC-1 was created of a PLC of 16 (K-12) classmates of 5 years, including learners with previously observed aggressive behavior, and VLC-2 was created of 21 classmates of over 6 years and a group of 9 learners, members of VLC-1, who had displayed aggressive behavior. VLC-1 was used for 7 months as a place for communication and collaboration for its learners and 2 teachers with active and instructive role, including creation and sharing of artifacts (e.g., videos, documents, presentations, pictures). VLC-2 had as a specific goal the conducting of an educational cultural project for 4 months, on a cultural topic selected by the learners, with 2 teachers with active and instructive role and 10 groups of 3 learners each, one of them being a learner of VLC-1 with aggressive behavior. Each group searched for information and created artifacts about a subtopic. Hybrid learning was the selected method, allowing collaboration either in the PLC and activities (e.g., museum visits) or mainly through the VLC, since offline interaction had positive impact on bonding.

Data preprocessing (data anonymization, segmentation in periods, letters to lowercase, tokenization, n-grams, removal of stop words, stemming, pruning of low/high-frequency terms, and length filtering) was conducted with the RapidMiner Studio and concerned the communication data of both VLCs, resulting in a data set consisting of 500 dialogue segments for VLC-1, and a data set consisting of 83 dialogue segments for VLC-2. One internal (active participant) annotator and one external annotator annotated each segment as “bullying” (1) or “no bullying” (0); regarding VLC-2, the agreement rate was 100% on none bullying instances, and regarding VLC-1, the agreement rate was 89%. The annotators also evaluated (scale 0–20): (a) the dialogues; their relevance regarding the topic, their effectiveness of communication, their number of threads, their linguistic quality, and their

[19] A multi-set of words based on a simplified representation for NLP and Information Retrieval.

semantic complexity (or simplicity), and (b) the artifacts; their relevance regarding the topic, their complexity (or simplicity), their quality and design, and their aesthetics.

Qualitative results showed that: (a) the participation was less in VLC-2 compared to VLC-1, due to inner speech; an indicator of the solid structure and robustness of a community, (b) there was no aggressive behavior in VLC-2, while in VLC-1 it was at a 15.4% rate, (c) teachers participation was at a 14% rate in VLC-1, and 43% in VLC-2, (d) regarding the artifacts, the mean value for VLC-1 was 4.25 and 15.82 for VLC-2, (e) low grades were observed for both VLCs; VLC-1 dialogues were irrelevant to the topic, and VLC-2 had a small amount of dialogues, and f. fear of non-acceptance could be the motivation for bullies to modify their behavior. ML experiments concerned VLC-1 (since VLC-2 had no bullying segments), and were conducted with Naive Bayes, Naive Bayes kernel, ID3, Decision Tree, feed-forward neural network, rule induction, and gradient boosted trees. The performance of all models was high; accuracy valid from 86.2% to 94.2%, precision from 88.65% to 99.53%, recall from 86.27% to 95.42%, and F1 score from 91.91% to 96.66%.

The effect of project-based activities on the aggressive behavior of the learners in VLCs was explored by the same authors in Nikiforos et al. [16]. Two PLCs (different from those of [4]) were transformed into VLCs in the Wikispaces web-based collaborative platform. The first VLC consisted of 21 schoolmates of 6 years and 1 teacher, and the platform was used for communication and sharing of the artifacts. The second VLC consisted of 21 schoolmates of 5 years, and the platform was used to conduct a project-based activity, in a topic selected from the learners. Groups of 3 to 4 learners collaborated and implemented a subtopic, by searching and collecting data, and creating and uploading artifacts. Both the dialogues and artifacts from the log files were analyzed, indicating that the absence of specific targeted activities had significant impact on aggressive behavior

Preprocessing included data anonymization and segmentation into periods, resulting in a data set consisting of 126 dialogue segments for the first VLC, and a data set consisting of 1167 segments for the second VLC. They were authentic humanistic data collected under real conditions. One internal (active participant) annotator and one external annotator annotated each segment as “bullying” (1) or “no bullying” (0); for the first VLC, the inter-annotator agreement was 92%, while 9.5% of the dialogues were considered to be bullying, and for the second VLC, the inter-annotator agreement was 98%, while 3% of the dialogues were considered to be bullying. The annotators also evaluated (scale 0–20) the dialogues, regarding the relevance to the topic and the communication effectiveness (3.11 for the first VLC and 13.25 for the second VLC), and the artifacts, regarding their relevance to the topic and their aesthetics (6.85 for the first VLC and 12.98 for the second VLC). Teachers participation was at a 5% rate in the first VLC, and 28% in the second VLC.

A more qualitative approach was explored in another work by Tzanavaris et al. [17], including discourse and artifacts analysis in VLCs to assess and outline the collaboration of learners. The specific characteristics of the dialogues in VLCs were considered. The VLC consisted of 20 (K-12) learners and 2 teachers with active and instructive role. Google Docs was used to collaborate in creating presentations for a cultural activity, as well as to discuss. Nine groups, of 2–3 learners each, were formed. The process was evaluated through individual questionnaires and interviews. Activity log files, dialogue text, questionnaires and interviews constituted the data that were analyzed, based on the Struggle Analysis Framework (both quantitative and qualitative), and the history of the presentations and analysis of dialogues. Semantic segmentation and annotation on the dialogues were applied. The researchers conducted action analysis, interaction analysis and evaluation of the presentations and the characteristics of the dialogues.

The data set created and used in the papers of this subsection is available for research purposes[20].

[20] https://inventory.clarin.gr/resources/browse/greek-student-chat-dataset/244fed8001c111eb867e00155d02a400f0da09cfeca14ba08781af781f78cf75/?fbclid=IwAR2uCLpqxnLuA2R67P8gvm_34C8yGfAjdHbNeKI3oZLutXXRQ-6EYtMIcQw, (accessed on 19 January 2021).

2.2.2. Offensive Language on Twitter

Pitenis et al. [6] created the first, manually annotated, data set in Modern Greek for offensive language identification (the Offensive Greek Tweet Dataset [21]. It consisted of 4779 tweets, which were collected from May to June 2019 and annotated as offensive and not offensive. The topics varied from television programs to elections. Several known Greek curse words and expletives were used as keywords for the collection of the data. Additionally, expressions structured with the auxiliary verb "to be" followed by an adjective or a noun were also considered. During preprocessing, Uniform Resource Locators (URLs) and emoticons were removed, while at-mentions, accentuation and duplicate punctuation were normalized. All letters were turned to lowercase. During annotation, the data set was tagged as offensive, not offensive, and spam, according to specific instructions. The inter-annotator agreement was considered for the final tags. The feature set consisted of TF-IDF n-grams and POS tags, and word embeddings, while Long short-term memory (LSTM) were used for the deep learning models. Experiments on the proposed data set were conducted with different subsets of the feature set and several baseline models: SVM, Stochastic Gradient Descent and Naive Bayes, as well as 6 deep learning models. LSTM and GRU with Attention obtained the best results, achieving an F1 score up to 89%.

Pontiki et al. [18] proposed a framework for verbal aggression analysis to study verbal attacks against specific target groups in Twitter, in order to identify and indicate xenophobic behaviors during the financial crisis in Greece (from 2013 to 2016). Verbal aggression was examined again 3 years later, for detecting changes on the target groups, the genre and content of the verbal attacks against the same groups during the refugee crisis following the 2019 elections. The results highlighted changes regarding the target groups of the verbal attacks, as well as the types and content of the attacks, in accordance with the perceptions and stereotypes during the financial crisis. Ten target groups were selected for examination: (a) Pakistani (TG1), (b) Albanians (TG2), (c) Romanians (TG3), (d) Syrians (TG4), (e) Muslims (TG5), (f) Jews (TG6), (g) Germans (TG7), (h) Roma (TG8), (i) Immigrants (TG9), and (j) Refugees (TG10). The data set consisted of 4,490,572 tweets, collected from 2013 to 2016, based on a set of keywords for each target group. Knowledge representation, computational analysis, and data visualization were applied. During preprocessing, tokenization, sentence splitting, POS tagging, and lemmatization were performed. The results for the period from 2013 to 2016 showed that the target groups of the most verbal attacks were TG6, TG2, TG1, TG5 and TG9. Most aggressive messages were rather directed to specific target groups, than expressing general and vague aggressive intentions. The stereotypes and prejudices were also detected from word frequencies, as represented in word clouds, and concerned mostly attacked target groups of this period. The results for the period following the 2019 elections showed that the target groups of the most verbal attacks were the same but in different order, the most aggressive messages concerned mainly the target group, and the stereotypes and prejudices concerned mostly attacked target groups. The researchers made some suggestions for future work: (a) extending the framework to other types of attacks, and (b) including other languages for cross-country and cross-cultural comparisons.

[21] https://sites.google.com/site/offensevalsharedtask/home, (accessed on 19 January 2021).

Table 3. Overview of the literature (offensive behavior and language detection). Social media, data sets and corpora, methods applied on data, and the resulting tool.

Paper	Social Media	Data	Methods	Tool
[4]	VLCs, Wikispaces	500 dialogue segments (VLC-1) 83 dialogue segments (VLC-2)	anonymization, segmentation in periods, manual annotation, lowercase, tokenization, n-grams, removal: stop words, stemming, pruning of low/high-frequency terms, length filtering	Detection of bullying behavior
[16]	VLCs, Wikispaces	126 dialogue segments 1167 dialogue segments	anonymization, segmentation in periods	Detection of bullying behavior
[17]	VLCs, Google Docs	activity log files, dialogue text, questionnaires, interviews	semantic segmentation, annotation	Discourse & artifacts analysis
[6]	Twitter	4779 tweets (May–June 2019)	keyword search, removal: emoticons, URLs, accentuation, normalization, lowercase, manual annotation, TF-IDF, n-grams, POS tags, word embeddings, LSTM	-
[18]	Twitter	4,490,572 tweets (2013–2016)	keyword search, knowledge representation, computational analysis, data visualization, tokenization, sentence splitting, POS tagging, lemmatization	-
[18]	Twitter	4,490,572 tweets (2013–2016)	keyword search, knowledge representation, computational analysis, data visualization, tokenization, sentence splitting, POS tagging, lemmatization	-

Table 4. Overview of the literature (offensive behavior and language detection). Machine learning and other algorithms, experimental results, contribution, and open issues.

Paper	Algorithms	Results	Contribution	Open Issues
[4]	Naive Bayes, Naive Bayes Kernel, ID3, Decision Tree, Feed-forward NN, Rule induction, Gradient boosted trees	accuracy up to 94.2%	1st study of the influence of VLCs on behavior modification regarding bullying NLP & ML framework for automatic detection of aggressive behavior & bullying authentic humanistic data collected under real conditions	-
[16]	Text analysis & annotation *t*-test	-	authentic humanistic data collected under real conditions	-
[17]	Struggle Analysis Framework	-	collaboration assessment action analysis interaction analysis evaluation of presentations & dialogues	-
[15]	SVM Stochastic Gradient Descent Naive Bayes 6 deep learning models	F1 score 89%	1st Greek annotated data set for offensive language identification	-
[18]	-	-	framework for verbal aggression analysis verbal attacks against target groups xenophobic attitudes during the Greek financial crisis	extending to other types of attacks including other languages for cross-country & cross-cultural comparisons

3. Opinion-Mining

Taking this work a step further, we focus on a quite well-known fact: millions of content creators worldwide produce a wealth of unstructured opinion data that exist online obtainable through popular crawling methods (i.e., Scrapy[22]) or through readily available platforms[23], while being generated when people share their opinions on several things, such as consumer experience. In principle, the intention to comment is voluntary, as it provides an honest view and opinion on a particular topic. Under this notion, the term of *opinion-mining* arises, since the analysis and summarization of large-scale data has led to a specific type of concept-based analysis [38]. In general, understanding public sentiment is the core action of implementing opinion-mining. There are many useful sources on the web, probably describing present opinion on politics, social matters, user reviews and many more, which are easily minable. On the other hand, it remains true that this novelty provides a volunteered source of highly esteemed user opinion. Although people express positive or negative feelings on a given topic (sentiment analysis), researchers need to understand the reasoning behind a given sentiment (opinion-mining); therefore, individual opinions are often reflective of a broader view. Given the large minable data sets, research groups need to develop new interpretation methods with the help of AI, to extract opinion from textual data. Nevertheless, such large data sets produce complex tasks that require arduous and tedious work on behalf of data scientists. Applying mining techniques for identifying the sentiment on the social web. Initially, texts are collected in the form of raw data and then they are preprocessed into specific data sets through ML and NLP approaches. Afterwards, researchers deploy various types of ML algorithms to detect web sentiment among a specific data set under the scope of analytical interpretation and assessment of the methodology in place. Recent research work has indicated that Greek social media presents a platform for users to express their opinion related to many aspects of private and social life and their experience with services and products. This section presents recent literature on the political footprint along with voting patterns (Section 3.1 [8,14,39–44]) and introduces to the reader work related to Marketing and Business Analysis (Section 3.2 [3,5]) that employ state-of-the-art opinion-mining ML techniques.

3.1. Politics and Voting Analysis

Greece has witnessed major political events during the last decade and subsequently Greek citizens, and voters in particular, are very often forced to reflect on their political preference based on broader occasions [45]. On that notion, there were many attempts to recognize the underlying patterns of social events by multidisciplinary scientific communities. The aim of this section is to explore whether the Greek media and social media discourses can provide discursive reconstruction on politics through state-of-the-art analytical methods. Table 5 presents a summary of works related to Greek text mining on Politics and Voting Analysis, which are discussed in this subsection.

[22] https://scrapy.org/, (accessed on 15 November 2020).

[23] http://insideairbnb.com/,https://www.pewresearch.org/download-datasets/, https://guides.library.cornell.edu/polling_survey_online, www.tripadvisor.com.gr, (accessed on 15 November 2020).

Table 5. Overview of the literature. Opinion-mining on Politics and Voting Analysis.

Paper	Social Media	Data	Methods	Tool	Algorithms	Results	Contribution	Open Issues
[39]	Twitter	57,424 tweets (April to May 2012)	sentiment analysis TF distribution	-	-	-	confirmation of the alignment between actual and social web-based political sentiment	implementation of more sophisticated text analysis techniques
[41]	Twitter	61.427 tweets (May 2012) divided into Parties & Leaders 44.438 tweets (after cleanup)	text classification, semantic analysis	OMW	NLTK	precision 82.4%	real-world application of irony detection	use of stemmer/lemmatizer, tool unavailability, small manually trained data set
[40]	Twitter	61,427 tweets (May 2012) divided into Parties & Leaders 44,438 tweets (after cleanup)	collective classification	OMW	J48, Naive Bayes, Functional Trees, K-Star, RF, SVM, Neural Networks	Supervised: Functional Trees 82.4% Semi-supervised: RF 83.1%	-	application with Word Vector or Deep Learning
[44]	Twitter	48,000 Tweets in two data sets (July & September 2015)	data collection and entity identification, volume analysis, entity co-occurrence, sentiment analysis and topic modeling	SentiStrength	-	highlight the societal and political trends	political domain analysis	bot recognition
[8]	Twitter	14,62M tweets, 283 Greek "stopwords"	convolutional kernels	User Voting intention modeling	SVM, LR, FF, RF	MCKL = 0.02%	real time systematic study on nowcasting the voting intention	annotating a random sample of Twitter users for increased performance
[42]	Twitter & Digital news media	540,989 articles (1996–2014)	PEA & NERC	NLP, NERC, EAU and FST	-	quantitative and qualitative	-	enrichment of sociopolitical event categories
[43]	Twitter & Digital news media	540,989 articles & 166,100,543 tweets (1996–2014)	PEA & NERC	NLP, NERC, EAU and FST	-	quantitative and qualitative	-	enrichment of sociopolitical event categories

One of the first complete approaches on Greek texts mining on political events was that of Kermanidis & Maragoudakis [39], where they propose a method for assessing political tweets before and after the election day focusing on the difference in web sentiment. This study indicated the degree of alignment between actual and social web-based political belief, related to electoral sentiment on major political events. The authors studied the impact of the acquired web sentiment before and after the Greek parliamentary elections of 2012 by implementing sentiment identification and Term Frequency (TF) distributions. Furthermore, this work negotiates the two-way alignment of actual political and web sentiment while using minimal linguistic resources.

A total of 57,424 tweets were exploited, dated from 29/04 to 13/05, 2012, and consequently they employed the names of major party leaders and parties as keywords on tweet collection. To obtain the most related tweets and gain some initial insight into the collected data, authors gained keywords in many different forms. The number of tweets was compared against the results of polls made by two large volume Greek newspapers, indicating alignment between tweet distribution and real political sentiment. Furthermore, each tweet set was tokenized according to a politician or political party. It should be noted that the use of a common stemmer is difficult due to various idioms such as "Greeklish", slang, word truncation and insufficient spelling check present on Greek tweets.

The authors decided to create a distinct vocabulary, contrary to common practice that dictates forming a vocabulary based on all tweet categories. Furthermore, polarity of sentiment was annotated manually due to lack of vocabulary words for each tweet. Then, each tweet set formed a separate data set and its corresponding vocabulary, according to specific politician or political party, in relation to election day. The authors proposed an unsupervised approach to tweet sentiment extraction under the notion of social media derived and real political sentiment among twitter users. Furthermore, their work introduced the implementation of advanced text analysis for sentiment extraction through co-occurrence statistics, which was based on tweet genre and language. Lastly, their work indicated the inability to detect significant pairs of terms for sentiment identification, learning techniques which remain relatively unexplored at this stage of the study.

Recently, a study from Charalampakis et al. [40,41] attempted to further categorize semi-supervised results from an updated version of their previous research. On the same data set[24], they implemented the collective classification technique, a semi-supervised learning method, which allows humorous political tweets to predict actual election results. Moreover, they deleted duplicate tweets to avoid frequency bias. The irony detection concept is based on subjective perceptions enabled by synsets[25], an embedded feature of Open Multilingual Wordnet (OMW)[26] lexical database [46]. In particular, this procedure allows to the formation of structural sentences and occurrences of unexpectedness, while detecting imbalance and unexpectedness, by categorizing them into five categories: Spoken, Rarity, Meanings, Lexical and Emoticons. All the above features produced low correlation (Pearson) among the variables. Training and testing were implemented using supervised and semi-supervised classifiers. The authors evaluated the algorithmic performance by employing the predicted data set through the following process: "train-set against the manual small data set", rendering it comparable among supervised and semi-supervised techniques. Furthermore, the data set was not directly correlated, but Charalampakis et al. revealed the correlations between ironic comments and partisan bias, as the ironic comments came mainly from exponents of the opposite view. However, there is not any clear differentiation in the performance of each algorithm, and it was observed that supervised classifiers produced more concrete results on ironic tweet detection. This work revealed reasonably acceptable performance of the aforementioned ML techniques and produced similar prediction scores on the fluctuation from past elections. Word vectors

[24] Available for research purposes upon request.

[25] Sets of cognitive synonyms.

[26] http://compling.hss.ntu.edu.sg/omw/, (accessed on 16 April 2021).

and deep learning could enhance the detection performance of this work, bridging the analysis with the election results.

Antonakaki et al. [44] deployed natural language analysis to two Twitter data sets. The analysis referred to a politically turbulent period in Greece (July and September 2015) triggered by a negotiating effort to restructure its national debt, a situation described by the first set that included all tweets that contained 301,000 tweets with the #*dimopsifisma* and #*greferendum* hashtags, and 182,000 tweets that contained the hashtags #*ekloges* and #*ekloges_round2*. The authors categorized their analysis into 5 distinct parts: data collection and Entity Identification, Volume Analysis, Entity co-occurrence, Sentiment Analysis and Topic Modeling. Focusing on Sentiment Analysis, they deployed a Greek language sentiment dictionary, embedded with the capability of detecting sarcasm through SentiStrength[27], a strength text detection tool for short texts. More particularly, it employs several methods to simultaneously extract each short informal text's strengths from each tweet and categorizing it into positive or negative sentiment. While recognizing the domain dependence of sentiment analysis, Antonakaki et al. created specialized lexicons by selecting and annotating existing words from known social web corpora. Consequently, they merged three Greek lexicons consisting of the existing SentiStrenth, the one implemented by SocialSensor[28] and the domain specific produced during this study. Sentiment polarity is the product of the aforementioned analysis that describes the duality of each tweet, expressing the intention to change the election result. Considering the particularities of Greek texts in social media ("Greeklish" and demographics) the authors succeeded in compiling sentiment and entity detection dictionaries for the Greek language specialized to political events. Furthermore, this study highlighted societal and political trends that enable public choices and actions, allowing the better understanding of public opinion and identification of the ways emotion drives societal and, consequently, political changes. Moreover, the capability of recognizing tweeting bots could improve the system's effectiveness.

Tsakalidis et al. [8] proposed a user voting intention model through a semi-supervised multiple convolution kernel learning approach. Their work consisted of an automated approach for Protest Event Extraction that allows the handling of sensitive text and network information from Twitter. Moreover, this study suggests a novel systematic approach on nowcasting voting intention while a political event is taking place, highlighting the incorporation of Human In The Loop—HITL through the implementation of multiple convolutional kernels. Their work was based on Greek 14.62 million tweets during the period of 2015 referendum that provided a set of approximately 283 common Greek stop words. It consisted of a developing platform that contained patterns of integration based on honest user opinion. The users' content and network presented temporal variations, indicated by both quantitative and qualitative analysis, and results that could indicate further improvements in fidelity and accuracy of their results.

Papanikolaou & Papageorgiou [42,43] proposed a data-driven and linguistically oriented framework for implementing protest analysis, based on the principles of Computational Social Science. In particular, protest analysis was studied through a Protest Event Analysis platform that handled digital media articles in Greece over the last two decades (1996–2014) through computer science. Their method exploited large sets of textual data obtained from the mass media, creating a Protest Event Database. In particular, the proposed platform is linguistically driven through the exploitation of many modules in each stage of the analysis. Morphosyntactic information from basic NLP workflow is a driver for techniques such as NLP, Event Analysis Unit (EAU) and Finite State Transducers-FST[29]. The authors proposed a semi-supervised methodology that initially recognizes structural elements of the event, links them together to complete the sets of events, and then records them through the Event Database. Their work consisted of a data-driven Event Extraction

[27] http://sentistrength.wlv.ac.uk/, (accessed on 16 April 2021).

[28] http://socialsensor.iti.gr/, (accessed on 16 April 2021).

[29] POS, lemmatization, chunking and parsing.

Methodology, which is segmented into five distinct steps: Events Coding, Data Collection, Data Exploration, Data Analysis and Data Visualization. Regarding the clear text analysis methodology, several written information types were involved, which can then be detected within the text and interlinked. This work consists of an automated interdisciplinary approach for Named Entity Recognition and Classification (NERC) divided into four major categories (Person, Organization, Location and Facility) for Protest Event Extraction from Greek texts while incorporating HITL. Papanikolaou & Papageorgiou recognized the need for further enrichment on sociopolitical categories in the given database, to produce better results. Moreover, the data set included tweets and news articles. They employed a quite simple and effective method for event extraction from Twitter that exploits a distinctive and commonly used feature, namely the hashtags filtered pool of tweets depending on the type of event (i.e., strikes). It is noteworthy that the extraction method from Twitter was semi-supervised, to detect the different information types that link the constituents to create an event tuple. Papanikolaou & Papageorgiou recognized the need for further enrichment on sociopolitical categories in the given database, to produce better results.

3.2. Marketing and Business Analysis

Sentiment analysis is an artificial intelligence technique that employs ML and NLP text analysis techniques to track polarity of opinion (positive to negative). A corporation, with the right tools, can gain insights from social media conversations, online reviews, emails, customer service tickets, and more. It has become an essential tool for marketing campaigns because it allows the researcher to automatically analyze data on a scale far beyond what manual human analysis could do, with unsurpassed accuracy, and in real time. Furthermore, it allows the approach of the mentality of a specific group of customers and the public at large to make data-driven decisions. More specifically, a corporation can even analyze customer sentiment and compare it against their competition, follow the emerging topics and check brand perception in new potential markets. The public offers millions of opinions about brands, services and products daily, on social media and within the world wide web. In Table 6 we present an overview related to literature on opinion-mining on Marketing and Business Analysis, which is discussed in this subsection.

This approach relies on ML and it is easier than other text analysis techniques to set up and manage. If the use of specific words or phrases used by the textual sample changes, inserting new contextual cues for the tool to work from can be difficult. The procedure of identifying and cataloging textual content according to subjective opinion on a product or/and service is characterized as Sentiment analysis [47]. Positive, negative and neutral feedback in the form of tweets or digital texts in general, is a useful tool for every business.

Table 6. Overview of the literature. Opinion-mining on Marketing and Business Analysis.

Paper	Social Media	Data	Methods	Tool	Algorithms	Results	Contribution	Open Issues
[5]	PaloPro	Blogs, Twitter and Facebook posts	sentiment analysis, reputation management, brand monitoring	OpinionBuster	NLP, CRFs	performance > 93%	sentiment and polarity detection of a word in its context	further optimization
[3]	SVM classifier	-	effectiveness of TF-IDF for automatic sentiment classifier for hotel reviews	Further use of contextual Valence shifters	SVM classifier	-	effectiveness of TF-IDF for automatic sentiment classifier for hotel reviews	further use of contextual Valence shifters

Within the Greek corpus in particular, Named Entity Recognition and Classification has seen its implementation into the field of reputation management. More specifically, Petasis et al. [5] deployed a brand monitoring service for the Greek language through an automated Software as a Service (SaaS) application that enables the large-scale linguistic and sentiment analysis for the Greek Web. This is offered by PaloPro[30], a polarity analysis platform, which ranges across different data inputs at the document and at the attribute level. More specifically, specific mentions on textual entities serves mining. As the first commercial automated platform for reputation management in Greece, it can handle heavy extraction duties through OpinionBuster, which incorporates the latest technical approaches to NLP, ranging from ontologies and rule-based systems based on rules to ML algorithms. Petasis et al. have developed an efficient platform capable of processing 100 documents per second. Subsequently it allows Named Entity Recognition tasks that locate mentions of entities related to a specific thematic domain. Moreover, OpinionBuster can perform tasks of recognizing specific objective statements that exploit a textual knowledge source with sentiment, on a specific context. The authors attempted to validate their system by manually annotating a corpus from two popular newspapers, taking into account all mentions of entities, while they also associated the polarity of these mentions with a specific time period. OpinionBuster achieved an accuracy of 80%, which was characterized as good performance. Their work presents a high-performance model (93%) regarding NLP implementation, which is mainly responsible for reporting entities and their antinomy for a specific domain in the Greek language. On the contrary, polarity detection did not exhibit the same performance levels (64%), indicating the need for further optimization.

Markopoulos et al. [3] built a sentiment-based system for classifying reviews through ML techniques. Their work focused on the hotel domain and included two different classification methods for documents referring to their overall sentiment of its content. This paper indicated an application of a ML approach which potentially presents better accuracy than semantic orientation approaches. Furthermore, Markopoulos et al. deployed a corpus of 1800 reviews consisting of hotel reviews from the Greek edition of Tripadvisor[31]. The data set did not include Greek translations of the review text due to many grammatical and syntactic errors, as well as extremely short or lengthy texts. At this point the authors maintained a balanced typology through equal selection of destinations and accommodation along the data set. Binary SVM classifiers have employed textual content as feature vectors by implementing feature selection on each word (uni-grams). More specifically, they applied the method dictated by the bag-of-words language model[32], and every individual is considered as a single word feature. Each feature cumulatively forms the data set, which represents the corpus consisting of a document per column matrix. This leads to a weighted model implemented by the methods of TF, TF-IDF, Term Occurrences (TO) and Binary Term Occurrences (BTO), consisting of the first features with the higher TF-IDF weight, and the positive or negative sentiment on unclassified documents through polarity prediction. Furthermore, this work compared two different classification methods[33], indicating the effectiveness of the TF-IDF method. It is worth mentioning that the authors cross-validated the performance of the classifiers at hand using accuracy, recall and precision as measures of positivity and negativity exclusively from Greek texts. Their prototype system is exploitable for providing useful information that could benefit both visitors and hotel owners through Tripadvisor or any other similar platform that supports review comments in the Greek language.

4. Discussion

Approaches on developing Modern Greek social web text data sets and on Modern Greek web text mining have been increasing in number in the last decade. The targeted

[30] https://palopro.io/en/, (accessed on 15 January 2021).

[31] www.tripadvisor.com.gr, (accessed on 15 January 2021).

[32] A text is represented as the total of the contained words.

[33] TF-IDF bag-of-words and TO.

social networks focus mainly on Twitter and to a less extent to Facebook data, mainly due to the easiness of collecting the corresponding data, as well as the researcher-friendly attitude of the first social network through the provision of well-documented APIs. Regarding bullying and aggressive behavior and language detection, most works focused on VLCs on Wikispaces and Google Docs collaborative platforms. The resulting data sets vary from data sets that are small in size (mainly due to the high level of manually generated linguistic annotations they incorporate, such as POS tags, ratings, and markups), to large, automatically generated and annotated, collections of Modern Greek social text. Mining applications are plentiful, and include POS taggers, sentiment and discourse analysis tools, bidialectal classifiers, authorship attribution and author's gender identification tools, argument extraction tools, and bullying and offensive language detectors.

Regarding the idiosyncrasies of the specific language, most approaches do not rely on language-specific lexica or other language-dependent tools. However, Modern Greek still presents several challenges, the most significant of which involve coping with the distinct alphabets, addressing the Greeklish writing, the multiple grammatical forms of the same word root that lead to large vocabulary sizes, and accentuation issues, as the meaning of the same word may change once its accent mark is moved onto another syllable. Additionally, the existence of dialects and idioms, such as Cretan, Cypriot and Pontic Greek, and their common use add to the complexity of the language. The limited linguistic processing resources, data sets and NLP tools pose more challenges for the researchers.

The contribution of the works analyzed in our paper is significant for text mining and NLP tasks in Modern Greek social text. Regarding linguistic patterns analysis [2,7,10–15,37], several data sets and corpora were created, a POS tagger, a POS tag set and a bidialectal classifier were introduced, several methods for sentiment intensity analysis, machine translation engines, argument extraction authorship and gender attribution were presented, and the most determinant features for bidialectal classifying, argument extraction and authorship and gender attribution were extracted.

Regarding behavioral patterns analysis (offensive behavior and language detection) [4,6,16–18], the influence of VLCs on behavior modification regarding bullying and aggressive behavior, as well as collaboration and interaction (dialogues and artifacts) analysis in VLCs were explored through NLP and ML frameworks for the detection of such behavior and language. Moreover, several data sets and corpora were created for both VLCs and Twitter, the latter mostly concerning xenophobic attacks against specific target groups.

Regarding politics and voting analysis, public opinion has been in the centerpiece of recent research works [8,39–44,48]. We have found a plethora of related studies attempting to tackle the traditional analysis of the political domain and social behavior. Mainly researchers went over blending entity-level sentiment and data statistics to assess the duality of public opinion. Twitter was identified to be the most popular social media platform, possibly due to the fact that is characterized by short texts and immediacy. Finally, an important point to consider that hinders the research process is the fact that Greek text is not always represented by the Greek alphabet and phraseology. On the contrary, almost every study had to decipher Greeklish and idioms to accurately extract the sentiment from each tweet.

Behavioral patterns related to consumer, market and business decision making is a complex procedure that involves many complex decisions, like comparing, evaluating, selecting, as well as synthesizing from a variety of services depending upon the opinion of a consumer over a particular product or commercial trend [49]. There is a small number of works [3,5] that studied the overall sentiment of each thematic domains for Marketing and business analysis. On the one hand, harvested opinion on marketing analysis was implemented cumulatively through a large-scale online platform that performs Named Entity Recognition tasks that allows continuous opinion monitoring on a specific company, organization, services or products. Furthermore, automatic sentiment classification was developed by comparing two different algorithmic methods (TF-IDF and TO) on specific "hand-picked" word lists. This allows implementing NLP for express positive or negative sentiment extraction of hotel reviews.

Finally, it should be noted that, at the time of writing this paper, Modern Greek data sets and text mining approaches are certainly fewer in number when compared against works focusing on other, more widely spoken, popular languages, such as English, German, Spanish, French, Italian, Russian or Chinese. In general, when compared to corresponding tasks in other languages, the obtained accuracy results in Modern Greek web text mining tasks indicate a better overall performance than [50–57], comparable to [58–60], or worse than [61–63].

5. Conclusions

In this survey paper we made an attempt to report on the basic stages and methodologies of indicative solutions that span across various aspects of mining techniques and applications on social web text that target Modern Greek, and as a result cover quite diverse research directions. More specifically, we have presented a survey on social web text data sets and text mining applications that have targeted the Modern Greek language. The challenges posed by the idiosyncrasies of the specific morphologically rich language have been described, and their effect on the mining accuracy has been reported.

From the above presentation and analysis, it becomes apparent that most methodologies and scenarios are based on state-of-the-art Data Mining, ML and NLP techniques. Overall, an obvious trend is to be identified and this may be summarized into the active use of ML methodologies and techniques for numerous NLP tasks; POS tagging, bidialectal classification, argument extraction, sentiment analysis, machine translation, authorship and gender attribution, and aggressive language detection, as well as Entity Identification, Volume Analysis, Entity co-occurrence, Collective classification, Sentiment Analysis and Topic Modeling.

It remains also quite interesting that the works taken into consideration and analyzed within our paper highlighted several open issues for text mining and NLP tasks in Modern Greek social text. Regarding linguistic patterns analysis [2,7,10–15,37], several researchers suggested experiments with larger data sets and feature sets, as well as more sophisticated ML algorithms, including data from more than one social media corpora, and extensions for Greeklish and more Greek dialects and variations. Moreover, they highlighted the need for more syntactic and semantic tools for Modern Greek, to implement applications in mapping linguistic diversity by region, tracking controversial events and their connections with users, and moderating of social media. Regarding behavioral patterns analysis (offensive behavior and language detection) [4,6,16–18], the researchers suggested extending their frameworks to other types of attacks, and including other languages for cross-country and cross-cultural comparisons.

The developments that have taken place in Greek political life have steered the interest of research community on recognizing political opinion while assessing the idiosyncrasy of the Modern Greek language. This work acknowledged several studies that have attempted to capture the impact of political and social developments over the last decade [8,14,39–44]. Regarding extracting political sentiment from textual data, researchers focused on analysis before and after the events (elections, referendum, etc.) based on large data sets from Twitter and Greek digital media. A considerable number of tools and techniques were materialized to be used for the computational implementation and analysis of the bulk of data collected. The proposed analytical processes led to various methodologies of associating intention on the web with real life political sentiment.

On the other hand, Business and Marketing analysis perceived as a method of identifying patterns and trends in relation to a particular product or service. The fact that social media is integrated into commercial applications and platforms now offers a large amount of data that is constantly increasing. It was observed that the recent literature on commercial applications has been removed from the analysis exclusively of the Greek language. We focused on certain studies [3,5] concerning the Greek language that served opinion-mining assessment on user experience as an automatic process of sentiment identification and standardization in text on large data sets from the Greek web and integrated social media. Both research approaches produced relatively accurate results on monitoring the observed sentiment in relation to specific services or products.

Thus, the main conclusion drawn from this work is that several Data Mining and NLP approaches have been developed for Modern Greek social text with overall high performance compared to similar approaches for other languages (both for high- and low-resourced ones). Tasks such as argument extraction and authorship and gender attribution have a clear potential for improvement. For future work, we consider examining innovative aspects of the Greek social web texts in narrower and trending application domains, such as emerging social networks (e.g., Tik Tok), to enable monitoring of the emerging trends in data use, as well as to identify potential novel data models. It is certainly not possible to address all aspects of modern data mining in a single paper nor discuss all expressions in the process. The interested reader will definitely identify a variety of open remaining issues, especially considering not only the researchers', but also the developers' point of view on the matter. As such, this paper hopes to assist both in taking this survey's observations into account within their future endeavors.

Author Contributions: A.D. did the preliminary basic research; M.N.N. and Y.V. contributed to the research and categorization of works and did the bibliographic search; P.M. and K.L.K. provided guidance and oversight. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
BILOU	Beginning-Inside-Outside-Unit
BTO	Binary Term Occurrences
CRF	Conditional Random Fields
EAU	Event Analysis Unit
FF	Feed-Forward neural network
FN	False Negative
FP	False Positive
FST	Finite State Transducers
HITL	Human In The Loop
HTML	Hypertext Markup Language
kNN	k-Nearest Neighbor
LR	Logistic Regression
LSTM	Long short-term memory
MCKL	Multiple Convolution Kernel Learning
ML	Machine Learning
MOOC	Massive Open Online Course
NERC	Name-Entity Recognition and Classification
NLP	Natural Language Processing
OMW	Open Multilingual Wordnet
PLC	Physical Learning Community
POS	Part-of-speech
PPV	Positive Predictive Value
RF	Random Forest
SaaS	Software as a Service
SVM	Support Vector Machine

TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TNR	True negative Rate
TP	True Positive
TO	Term Occurrences
URL	Uniform Resource Locator
UTF	Unicode Transformation Format
VLC	Virtual Learning Community
WEKA	Waikato Environment for Knowledge Analysis

References

1. Alexandridis, G.; Michalakis, K.; Aliprantis, J.; Polydoras, P.; Tsantilas, P.; Caridakis, G. A Deep Learning Approach to Aspect-Based Sentiment Prediction. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Neos Marmaras, Greece, 5–7 June 2020; Springer: Cham, Switzerland, 2020; pp. 397–408.
2. Nikiforos, M.N.; Kermanidis, K.L. A Supervised Part-Of-Speech Tagger for the Greek Language of the Social Web. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 3861–3867.
3. Markopoulos, G.; Mikros, G.; Iliadi, A.; Liontos, M. Sentiment analysis of hotel reviews in Greek: A comparison of unigram features. In *Cultural Tourism in a Digital Era*; Springer: Cham, Switzerland, 2015; pp. 373–383.
4. Nikiforos, S.; Tzanavaris, S.; Kermanidis, K.L. Virtual learning communities (VLCs) rethinking: Influence on behavior modification—Bullying detection through machine learning and natural language processing. *J. Comput. Educ.* **2020**, *7*, 531–551. [CrossRef]
5. Petasis, G.; Spiliotopoulos, D.; Tsirakis, N.; Tsantilas, P. Sentiment analysis for reputation management: Mining the greek web. In Proceedings of the Hellenic Conference on Artificial Intelligence, Ioannina, Greece, 15–17 May 2014; Springer: Cham, Switzerland, 2014; pp. 327–340.
6. Pitenis, Z.; Zampieri, M.; Ranasinghe, T. Offensive language identification in greek. *arXiv* **2020**, arXiv:2003.07459.
7. Sababa, H.; Stassopoulou, A. A classifier to distinguish between cypriot greek and standard modern greek. In Proceedings of the 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 15–18 October 2018; pp. 251–255.
8. Tsakalidis, A.; Aletras, N.; Cristea, A.I.; Liakata, M. Nowcasting the stance of social media users in a sudden vote: The case of the Greek Referendum. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Turin, Italy, 22–26 October 2018; pp. 367–376.
9. Vallet, D.; Fernandez, M.; Castells, P.; Mylonas, P.; Avrithis, Y. A contextual personalization approach based on ontological knowledge. In Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 2006), Contexts and Ontologies: Theory, Practice and Applications, Riva del Garda, Italy, 28 August–1 September 2006
10. Mikros, G.K. Authorship attribution and gender identification in Greek blogs. *Methods Appl. Quant. Linguist.* **2012**, *21*, 21–32.
11. Baxevanakis, S.; Gavras, S.; Mouratidis, D.; Kermanidis, K.L. A machine learning approach for gender identification of Greek tweet authors. In Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 30 June–3 July 2020; pp. 1–4.
12. Kalamatianos, G.; Mallis, D.; Symeonidis, S.; Arampatzis, A. Sentiment analysis of Greek tweets and hashtags using a sentiment lexicon. In Proceedings of the 19th Panhellenic Conference on Informatics, Athens, Greece, 1–3 October 2015; pp. 63–68.
13. Goudas, T.; Louizos, C.; Petasis, G.; Karkaletsis, V. Argument extraction from news, blogs, and social media. In *Hellenic Conference on Artificial Intelligence*; Springer: Cham, Switzerland; Ioannina, Greece, 15–17 May 2014; pp. 287–299.
14. Goudas, T.; Louizos, C.; Petasis, G.; Karkaletsis, V. Argument extraction from news, blogs, and the social web. *Int. J. Artif. Intell. Tools* **2015**, *24*, 1540024. [CrossRef]
15. Sardianos, C.; Katakis, I.M.; Petasis, G.; Karkaletsis, V. Argument extraction from news. In Proceedings of the 2nd Workshop on Argumentation Mining, Lisbon, Portugal, 17–21 September 2015; pp. 56–66.
16. Nikiforos, S.; Tzanavaris, S.; Kermanidis, K.L. Bullying Behavior and Project-based Activities in Virtual Learning Communities (VLCs). In Proceedings of the 2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Corfu, Greece, 25–27 September 2020; pp. 1–5.
17. Tzanavaris, S.; Nikiforos, S.; Mouratidis, D.; Kermanidis, K.L. Virtual Learning Communities (VLCs) rethinking: From negotiation and conflict to prompting and inspiring. *Educ. Inf. Technol.* **2020**, *26*, 257–278. [CrossRef]
18. Pontiki, M.; Gavriilidou, M.; Gkoumas, D.; Piperidis, S. Verbal Aggression as an Indicator of Xenophobic Attitudes in Greek Twitter during and after the Financial Crisis. In Proceedings of the Workshop about Language Resources for the SSH Cloud, Marseille, France, 11–16 May 2020; pp. 19–26.
19. Lo, S.L.; Cambria, E.; Chiong, R.; Cornforth, D. Multilingual sentiment analysis: From formal to informal and scarce resource languages. *Artif. Intell. Rev.* **2017**, *48*, 499–527. [CrossRef]

20. Cambria, E.; Das, D.; Bandyopadhyay, S.; Feraco, A. Affective computing and sentiment analysis. In *A Practical Guide to Sentiment Analysis*; Springer: Cham, Switzerland, 2017; pp. 1–10.
21. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2020; pp. 1–3,11,12.
22. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2003; pp. 798–801, 852, 853.
23. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
24. Montague, P.R. Reinforcement learning: An introduction, by Sutton, RS and Barto, AG. *Trends Cogn. Sci.* **1999**, *3*, 360. [CrossRef]
25. Van Otterlo, M.; Wiering, M. Reinforcement learning and markov decision processes. In *Reinforcement Learning*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 3–42.
26. Petasis, G.; Karkaletsis, V.; Paliouras, G.; Androutsopoulos, I.; Spyropoulos, C.D. Ellogon: A new text engineering platform. *arXiv* **2002**, arXiv:cs/0205017.
27. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the European Conference on Information Retrieval, Santiago de Compostela, Spain, 21–23 March 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 345–359.
28. Thanopoulos, A.; Kermanidis, K.; Fakotakis, N. Challenges in extracting terminology from Modern Greek texts. In Proceedings of the 3rd International Workshop on Text-Based Information Retrieval (TIR-06), Riva del Garda, Italy, 28 August–1 September 2006; p. 53.
29. Clackson, J. *Indo-European Linguistics: An Introduction*; Cambridge University Press: Cambridge, UK, 2007.
30. Barðdal, J.; Smitherman, T.; Bjarnadóttir, V.; Danesi, S.; Jenset, G.B.; McGillivray, B. Reconstructing constructional semantics: The dative subject construction in old norse-icelandic, latin, ancient greek, old russian and old lithuanian. *Stud. Lang. Int. J. Spons. Found. Found. Lang.* **2012**, *36*, 511–547.
31. Sido, J.; Pražák, O.; Přibáň, P.; Pašek, J.; Seják, M.; Konopík, M. Czert–Czech BERT-like Model for Language Representation. *arXiv* **2021**, arXiv:2103.13031.
32. Husain, F.; Uzuner, O. A Survey of Offensive Language Detection for the Arabic Language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2021**, *20*, 1–44. [CrossRef]
33. Lopez, C.E.; Vasu, M.; Gallemore, C. Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. *arXiv* **2020**, arXiv:2003.10359.
34. Vilares, D.; Peng, H.; Satapathy, R.; Cambria, E. BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1292–1298.
35. Athanasiou, V.; Maragoudakis, M. A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: A case study for modern Greek. *Algorithms* **2017**, *10*, 34. [CrossRef]
36. Chatzikyriakidis, S. Clitics in Four Dialects of Modern Greek: A Dynamic Account. Ph.D Thesis, University of London, London, UK, 2010.
37. Sosoni, V.; Kermanidis, K.L.; Stasimioti, M.; Naskos, T.; Takoulidou, E.; Van Zaanen, M.; Castilho, S.; Georgakopoulou, P.; Kordoni, V.; Egg, M. Translation crowdsourcing: Creating a multilingual corpus of online educational content. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
38. Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [CrossRef]
39. Kermanidis, K.L.; Maragoudakis, M. Political sentiment analysis of tweets before and after the Greek elections of May 2012. *Int. J. Soc. Netw. Min.* **2013**, *1*, 298–317. [CrossRef]
40. Charalampakis, B.; Spathis, D.; Kouslis, E.; Kermanidis, K. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Eng. Appl. Artif. Intell.* **2016**, *51*, 50–57. [CrossRef]
41. Charalampakis, B.; Spathis, D.; Kouslis, E.; Kermanidis, K. Detecting irony on greek political tweets: A text mining approach. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS), Rhodes, Greece, 25–28 September 2015; pp. 1–5.
42. Papanikolaou, K.; Papageorgiou, H.; Papasarantopoulos, N.; Stathopoulou, T.; Papastefanatos, G. "Just the Facts" with PALOMAR: Detecting Protest Events in Media Outlets and Twitter. In Proceedings of the International AAAI Conference on Web and Social Media, Cologne, Germany, 17–20 May 2016; Volume 10.
43. Papanikolaou, K.; Papageorgiou, H. Protest Event Analysis: A Longitudinal Analysis for Greece. In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, Marseille, France, 11–16 May 2020; pp. 57–62.
44. Antonakaki, D.; Spiliotopoulos, D.; Samaras, C.V.; Pratikakis, P.; Ioannidis, S.; Fragopoulou, P. Social media analysis during political turbulence. *PLoS ONE* **2017**, *12*, e0186836. [CrossRef] [PubMed]
45. Tziovas, D. *Greece in Crisis: The Cultural Politics of Austerity*; Bloomsbury Publishing: London, UK, 2017.
46. Bond, F.; Fellbaum, C.; Hsieh, S.K.; Huang, C.R.; Pease, A.; Vossen, P. A multilingual lexico-semantic database and ontology. In *Towards the Multilingual Semantic Web*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 243–258.
47. Alessia, D.; Ferri, F.; Grifoni, P.; Guzzo, T. Approaches, tools and applications for sentiment analysis implementation. *Int. J. Comput. Appl.* **2015**, *125*.

48. Charalabidis, Y.; Loukis, E.N.; Androutsopoulou, A.; Karkaletsis, V.; Triantafillou, A. Passive crowdsourcing in government using social media. *Transform. Gov. People Process Policy* **2014**, *8*, 283–308. [CrossRef]
49. Ramaswamy, V.; Gatignon, H.; Reibstein, D.J. Competitive marketing behavior in industrial markets. *J. Mark.* **1994**, *58*, 45–55. [CrossRef]
50. Aldayel, H.K.; Azmi, A.M. Arabic tweets sentiment analysis–a hybrid scheme. *J. Inf. Sci.* **2016**, *42*, 782–797. [CrossRef]
51. Psomakelis, E.; Tserpes, K.; Anagnostopoulos, D.; Varvarigou, T. Comparing methods for twitter sentiment analysis. *arXiv* **2015**, arXiv:1505.02973.
52. Tripathi, P.; Vishwakarma, S.K.; Lala, A. Sentiment analysis of english tweets using rapid miner. In Proceedings of the 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 12–14 December 2015; pp. 668–672.
53. Shoemark, P.; Kirby, J.; Goldwater, S. Inducing a lexicon of sociolinguistic variables from code-mixed text. In Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, Brussels, Belgium, 1 November 2018; pp. 1–6.
54. Trye, D.; Calude, A.S.; Bravo-Marquez, F.; Keegan, T.T.A.G. Māori loanwords: A corpus of New Zealand English tweets. Presented at the Vocab@ Leuven 2019, Florence, Italy, 1–3 July 2019.
55. Erdmann, A.; Habash, N. Complementary strategies for low resourced morphological modeling. In Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology, Brussels, Belgium, 31 October 2018; pp. 54–65.
56. Foster, J.; Cetinoglu, O.; Wagner, J.; Le Roux, J.; Hogan, S.; Nivre, J.; Hogan, D.; Van Genabith, J. # hardtoparse: POS Tagging and Parsing the Twitterverse. In Proceedings of the AAAI-11 Workshop on Analyzing Microtext, San Francisco, CA, USA, 7–11 August 2011.
57. Bach, N.X.; Linh, N.D.; Phuong, T.M. An empirical study on POS tagging for Vietnamese social media text. *Comput. Speech Lang.* **2018**, *50*, 1–15. [CrossRef]
58. Öztürk, N.; Ayvaz, S. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telemat. Inform.* **2018**, *35*, 136–147. [CrossRef]
59. Carneiro, H.C.; França, F.M.; Lima, P.M. Multilingual part-of-speech tagging with weightless neural networks. *Neural Netw.* **2015**, *66*, 11–21. [CrossRef] [PubMed]
60. Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J.; Smith, N.A. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2010; pp. 42–47.
61. Gao, W.; Fang, Y.; Wang, Y.; Zhang, F. HRCE: Detecting Food Security Events in Social Media. *J. Phys. Conf. Ser.* **2020**, *1437*, 012090. [CrossRef]
62. Popescu, A.M.; Pennacchiotti, M. Detecting controversial events from twitter. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, 26–30 October 2010; pp. 1873–1876.
63. Popescu, A.M.; Pennacchiotti, M.; Paranjpe, D. Extracting events and event descriptions from twitter. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 105–106.