Modern Greek on Social Web

Subjects: Others | Computer Science, Artificial Intelligence | Computer Science, Interdisciplinary Applications
Contributor: Phivos Mylonas

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.

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

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 [1,2,3,4,5,6,7]. 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 [2,7,10,11,12,13,14,15] and behavioral [4,6,16,17,18] 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 [2,4,7,10,11,13,14,15,16,19,19]. 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].

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 [2,4,7,10,11,13,14,15,16]. 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.

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 [13,14,15]. 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 Table 1 and Table 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	tokenization, normalization,	POS tagger
		31,697 tokens (April 2019)	encoding, annotation	
[12]	Twitter	4,373,197 tweets	automated & manual rating,	Sentiment analysis lexicon
		30,778 users	removal: stop words & tone marks,	
		54,354 hashtags (April 2008–November 2014)	stemming, uppercase	
[7]	Twitter	1039 sentences	anonymization, manual annotation,	Bidialectal classifier
	Facebook	7026 words (Cypriot Greek)	removal: tabs, newlines, duplicate punctuation,	
	forums, blogs	7100 words (Modern Greek) (March–April 2018)	insertion: spaces, n-grams, encoding, tokenization	
[37]	MOOC	multilingual corpus	conversion into plain text,	-
		course forum text	removal: special characters, non-content lines,	
		quiz assessment text	multiple whitespaces, tokenization, sentence segmentation,	
		subtitles of online video lectures	special elements markup	
[13]	Twitter, news	204 documents	manual annotation, tokenization, sentence splitting,	-
	blogs, sites	16,000 sentences:	POS tagging, feature selection, gazetteer lists, lexica, TF-IDF	
		760 argumentative		
[14]	Twitter, news	204 documents	manual annotation, tokenization, sentence splitting,	-
	blogs, sites	16,000 sentences:	POS tagging, feature selection, gazetteer lists, lexica, TF-IDF	
		760 argumentative		
		comparison with NOMAD data set		
[15]	Twitter	1st: 77 million documents	POS tagging, cue words, distributed representations of words,	-
	Facebook	2nd: 300 news articles,	feature extraction, sentiment analysis, lowercase	
	news, blogs	1191 argumentative segments		

Paper	Social Media	Data	Methods	Tool
[10]	Blogs	1000 blog posts	stylometric variables, character & word uni-grams,	Authorship attribution &
		406,460 words (September 2010–August 2011)	bi-grams, tri-grams, feature extraction	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	accuracy up to 99.87%	1st data set for Greek social text	larger data sets
	ID3		1st tag set	data from different social media
			1st supervised POS tagger	syntactic & semantic analysis tools
				linguistic diversity by region
				tracking controversial events &
				mapping connections with users
[12]	Pearson Kendall	sentiment correlation	public benchmark data set	lexicon for social text
	correlation		set of intensity rated tweets	more linguistic data
			automated method for detecting	larger data set & number of raters
			intensity (tweets & hashtags)	
			temporal changes in intensity (hashtags)	
[7]	Naive Bayes, SVM, LR	95% mean accuracy	1st classifying Greek dialects in social text	applications in social media moderation
			bidialectal corpus & classifier	and academic research
			most informative features	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	more features & algorithms
			novel corpus	testing of Markov models

Paper	Algorithms	Results	Contribution	Open Issues
			most determinant features	
[14]	LR, RF, SVM, CRF	accuracy up to 77.4%	2-step argument extraction	more features & algorithms
			novel corpus	testing of Markov models
			most determinant features	comparing performance with approaches for English
				experiments with unsampled data
[15]	word2vec CRF	up to 39.7% precision	semi-supervised multi-domain method	extending the gazetteer lists
		27.59% recall	argument extraction	bootstrapping on CRF
		32.53% F1 score	novel corpus	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	more features combining gender & age
			NLP framework for gender identification of the author	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 Table 3 and Table 4.

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)	anonymization, segmentation in periods, manual annotation,	Detection of bullying behavior
		83 dialogue segments (VLC-2)	lowercase, tokenization, n-grams, removal: stop words,	
			stemming, pruning of low/high-frequency terms, length filtering	
[16]	VLCs, Wikispaces	126 dialogue segments	anonymization, segmentation in periods	Detection of bullying behavior
		1167 dialogue segments		
[17]	VLCs, Google Docs	activity log files, dialogue text,	semantic segmentation, annotation	Discourse & artifacts analysis
		questionnaires, interviews		
[6]	Twitter	4779 tweets	keyword search, removal: emoticons, URLs, accentuation,	-

Paper	Social Media	Data	Methods	Tool
		(May–June 2019)	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., Scrapy22) or through readily available platforms23, 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 [8,14,39,40,41,42,43,44]) 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.

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

Paper	Social Media	Data	Methods	Tool	Algorithms	Results	Contribution	Open Issues
[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.

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.

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