Social Media Prediction: A Literature Review

Dimitrios Rousidis, Paraskevas Koukaras, Christos Tjortjis*
School of Science and Technology, International Hellenic University
14th km Thessaloniki - Moudania
GR-570 01 Thermi, Greece
{d.rousidis, p.koukaras, c.tjortjis}@ihu.edu.gr

Abstract: Social Media Prediction (SMP) is an emerging powerful tool attracting the attention of researchers and practitioners alike. Despite its many merits, SMP has also several weaknesses, as it is limited by data issues, like bias and noise, and the lack of confident predictions and generalizable results. The goal of this paper is to survey popular and trending fields of SMP from 2015 and onwards and discuss the predictive models used. We elaborate on results found in the literature, while categorizing the forecasting attempts, based on specific values (source of data, algorithm used, outcome of prediction etc.). Finally, we present our findings and conduct statistical analysis on our dataset and critique the outcome of the attempted prediction reported by the reviewed papers. Our research indicates that results are ambiguous, as not all forecasting models can predict with high accuracy, and prediction seems dependable on the associated field, although some of the documented attempts are promising. More than half (53.1%) of the examined attempts achieved a valid prediction, nearly one fifth (18.8%) did not, while the remaining 28.1% is characterized as plausible or partially validated. By reviewing recent and up-to-date literature and by providing statistics, this paper provides SMP researchers with a guide on methods, algorithms, techniques, prediction success and challenges on three main categories that aid SMP exploration.

Keywords
Social Media Prediction, Data Prediction Models, Social Media Data, Web Information Systems, Web tools and applications

Abbreviations
SM (Social Media), SMP (Social Media Prediction)

Corresponding author
Assistant Professor Christos Tjortjis
ORCID iD: 0000-0001-8263-9024
Telephone: +30 2310807576

Email: c.tjortjis@ihu.edu.gr

1. Introduction

Social Media (SM) have recently established their dominance as multifunctional tools and many corporate colossuses, companies, institutes or even individuals organize and implement their strategies and promote and disseminate their businesses based on SM. Forecasting techniques lead the competition and play a vital role in surviving and growing in this ecosystem.

Nowadays, people around the world use SM to communicate, connect and interact with other users, becoming producers of Big Data at a great rate [18]. For a long time, the most common form of data was the traditional, mostly relational one, suitable for data mining techniques [19]. The past years new practices that embed, encrypt and process data with various practices, like steganography, "a technique of covert communication that aims to embed secret messages into an innocent carrier signal by slightly altering its most insignificant components, such that an unauthorized user will not be aware of the existence of secret data" [98] have been introduced and researched [96], [97], [98]. These practices have been used in SMP [103], [104], [105], making the volume of data even bigger and more difficult and at the same time intriguing to decode.

Social Media mining is a new, fast developing and growing field which should deal with noisy, free-format and sometimes long data or different types of multimedia [24], [25]. SM data might contain information about individuals, opinions, trends, pricings, reviews, health, incidents etc. Relations or links between entities or other social networks should be utilized, using social data along with statistical and mathematical methodologies harnessing SM data knowledge [102].

Many works discuss Social Media Prediction (SMP), opinion mining and information network mining techniques trying to establish standardizations regarding the *predictive power* and caveats of information contained in SM data [27], [3], [28], [29], [30], [62], [99], [100], [101], [106]. Opinion mining or sentiment analysis "*deals with analyzing people's opinions, sentiments, attitudes and emotions towards different brands, companies, products and even individuals*" [99]. In [100], Petz et al. discussed pre-processing approaches that aid the researchers with the issues of sentiment analysis in real world situations. In [99] and [101] the authors identified the differences of SM channels like microblogs, social network services, weblogs, discussion forums and product review sites and in [101] Petz et al. evaluated the effectiveness of several text preprocessing algorithms as a subtask of opinion mining in the aforementioned social media channels.

A major reason for studying SM forecasting techniques is the possibility of helping enterprises serve their customers better, through recognizing or predicting trends or governments prevent and tackle important matters before they even impact on communities.

In this paper we analyze up to date models (from 2015 up to 2019) and techniques that attempt to predict future outcomes and incidents in three fields (Finance, Marketing and Sociopolitical), and we examine their validity along with the algorithms and the methodologies they utilize, providing a survey of forecasting attempts. The main goals of our work are to provide a list of recent and related studies in SMP, to investigate prediction difficulties in some fields and to examine which Data Mining algorithms seem to work better and why. By doing so, we will be able to offer a survey that acts as a guide for other researchers. We recognize prediction methods that are effective (or not too effective) whilst highlighting areas for future improvements. In section 2 we present related work and propose domain categories for SM predictions. In section 3 we discuss SM prediction attempts forming the proposed categories. Section 4 presents a statistical analysis and appraises our findings. We conclude with general discussion on this topic and directions for further work in sections 5 & 6.

2. Related Work

Predictive analysis with SM is an emerging trend, which can be employed in numerous real-life applications [41], [54], [61]. Recently, researchers and academic communities deal with evolving matters and studies of computer and social sciences [31], [32]. Rapidly evolving literature is presented, showing a great interest and need for disciplines and new tools for handling SMP, aiming at producing highly acceptable *insights*. These insights can be very useful for other scientific fields.

SMP is used by a plethora of corporations, industries and organizations to enhance their business by predicting behaviors and trends [1], [2], [20]. More specifically, 65 of the top 100 Fortune-500 companies have utilized at least one of the four major Social Network Sites (Facebook 87.7% of these 65 companies, Twitter 84.6%, YouTube 56.9% and LinkedIn 32.3%) in 2013 [20]. Lassen et al. conducted a systematic literature review on predictive analysis with SM and they presented the issues in SMP modeling, along with the powerhouse of information that SM analytics provide [41]. Another major literature review has been conducted to study how SM can be used to predict the future [54].

Yu and Kak analyzed five (5) major subjects (*Marketing, Movie box office, Information dissemination, Elections and Microeconomics*) that researchers have based their studies to make accurate predictions using SM data [3]. The authors also categorized the predictors that were implemented in the literature (message and social network characteristics) along with their metrics (sentiment, time-series, terminology, degree, density, centrality, structural hole). Finally, they listed the prediction methods that were used in the literature (regression method, Bayes classifier, k-nearest neighbor classifier, artificial neural network, decision trees and model-based prediction).

In [29], Asur and Huberman introduced a taxonomy of prediction models and provided the corresponding literature for five (5) subfields: i) electoral results, ii) stock market movement, iii) iv) product sales and iv) influenza incidence. Issues and difficulties for forecasting models in SM were examined along with nowcasting, the handling of noise and biases and the comparison between aggregate and individual prediction. Finally, they concluded that statistical models are the most prolific methods to use in order to make valid predictions from SM data.

Lassen et al. examined the predictive analytics with SM data [41]. Initially, they discussed the difference between predictive and explanatory models along with the advantages of predictive models. They also presented the problems and the issues researchers face when using predictive modeling on SM data, they presented papers on SMP categorizing them by application domain, SM platform, variables and the statistical methods employed. They concluded with reflections on SMP.

Finally, a remarkable literature review was conducted by Philips et al. [54] in 2017. They provided a very extensive study by questioning how well SM data can be used to predict the future. They examined the topic, data source, data size, features, task and success rate of 107 articles on a time span of one decade, that were grouped in five (5) categories and twelve (12) subfields: i) Political Science (with 2 fields), ii) Economics (with 3 fields), iii) Public Health (with 3 fields), iv) Threat Detection (3 fields) and v) User Characteristics (2 fields). They concluded that "*SM forecasting is limited by data biases, noisy data, lack of generalizable results, a lack of domain-specific theory, and underlying complexity in many prediction tasks*". However, they were optimistic about the future of SMP that better and more accurate results will be achieved.

Having the aforementioned studies as a basis, our work combines research work reported in the literature and accumulates most of the features examined, like the algorithms used and also statistics that examine whether the SM used, the volume of data, the algorithms, the number of SM, the researched field have any effect on the outcome of the prediction model proposed.

More specifically, we provided new and up-to-date literature as [3] was published on 2012, [29] on 2013, [41] and [54] on 2017. We examined nine (9) fields (with one of them including health, diseases and vaccines) that were grouped in three (3) categories, whilst [3] and [29] provided related work for five (5) fields, [41] did not include any categories (although the authors gave emphasis on sales, entertainment, health, finance and politics with around five (5) papers per category) and [54] examined twelve (12) fields from five (5) categories. In addition, compared to the other four (4) related work studies, we and [54] are the only ones providing the prediction outcome of each method and only our study provided statistics. Finally, we added key-findings to a table Summarizing SMP attempts.

3. Social Media Prediction

There is a vast variety of fields that disperse widely and therefore in this paper we have grouped them in three (3) major categories to make their study easier.

- The first category is *Finance*, which ranges from public finance that revolves around government policies up to corporate finance, like stock markets and personal finance which is focused on family and household budgets, like predicting product pricing.
- The second category is *Marketing* and the needs of the industry various organizations to predict trends, behaviors and needs.
- The third category deals with *Sociopolitical* attempts for prediction and it includes fields like elections, diseases, natural phenomena, etc.

Data analysts have used a variety of variables, but in this survey, we have chosen to focus just on the SM ones. *Table 1* depicts the three (3) categories and their fields that we analyzed.

Table 1: Categories for SM Prediction

Finance	**Marketing**	**Sociopolitical**
Stock Markets	Customer Needs	Elections
Product Pricing	Trends & Entertainment	Health/Diseases/ Vaccines
Real Estate	Product Promotion	Natural Phenomena

Methodology

The methodology followed for identifying the papers for our survey was simple and straight forward. We studied papers that were published from 2015 and onwards as we wanted to have a recent and updated dataset and to examine the new methodologies that have been used by the researchers the past three (3) years in order to predict via SM.

Since there are numerous fields that can be examined, we decided to use the three (3) ones that as we studied the literature proved to be very popular.
We used Google Scholar to apply the date filter and to search the papers using combination of the major keywords such as “social media”, “social networks” and “prediction” or “forecasting”. Then we added keywords from the 3 fields “finance”, “marketing” and “sociopolitical” and from their subfields like Stock Markets, Product Pricing, “Real Estate”, “Customer Needs”, “Entertainment”, “Product Promotion”, “Elections”, “Health”, “Diseases”, “Vaccines”, “Natural Phenomena”. We selected the papers from the first five (5) pages of the results and when a large number of results for a specific field were appearing as we wanted a balanced dataset, we chose the most cited ones.

We managed to analyze forty-three (43) papers that met the criteria of our aforementioned methodology.

3.1 Finance

SM provide a powerhouse of data for prediction in finance and economy, especially if it is combined with other information from the web [8]. Kappri and Crawford investigated the damage made to the stocks of an enterprise due to a hacked tweet at the Associated Press Twitter account [5]. The authors stated that “In short, Twitter affects not only human traders but the whole stock market system structured with people, code and computers”.

The systematic literature review conducted by Schade echoed the results of many studies and indicated that it is feasible to use SM data to predict market needs and trends [40]. Finally, research used 5.7 million geotagged Tweets to provide a methodology to investigate and understand socio-spatial relations in big cities [26].

3.1.1 Stock Markets

SM opinions were examined if they can predict enterprises’ future credit risk with the aid of regression [4]; a strong connection (79.13% accuracy) was found between them, SM opinions can: “actually provide valuable information to retail investors”.

Dong et al. achieved 83.57% fraud prediction accuracy by combining SM features with SeekingAlpha and by utilizing Document Classification and Support Vector Machine (SVM) model with Gaussian Kernel [6].

Sun et al. provided a model that outclasses a baseline regression to predict the stock market by using text mining on a high-volume data from SM and by ignoring sentiment [7]. By proposing a Sparse Matrix factorization (SFM) model, they managed to achieve a prediction accuracy of 51.37% for daily prediction frequency.

A research was conducted about the price of Apple Inc. stock and the volume of the tweets about the enterprise [11]. The author implemented a Sentiment Classifier (SentiStrength). According to the results, there is no positive correlation of mood indicators and tweet volume with prices and there is a positive one between tweet volume and trading volume. In the same study, the author identified a negative correlation of search word volume with closing price and a positive one with volatility and trading volume.

Jit et al. used data from Google search, Bloomberg News and Twitter in their proposed Delta Naive Bayes (DNB) approach [17]. Their study showed that the weighted average of Precision and Recall measures can be as high as 0.407 when all the three sources are utilized, whilst the *F1* measure is as low as 0.147 when using two sources, and with a single source the measure is just 0.138.

A study about the implementation of sentiment analysis from Twitter Data in order to predict stock market movements showed that it is feasible to achieve a high percentage of accuracy [22]. The authors demonstrated that they could achieve a Sentiment Analysis accuracy of 70.5% with N-gram and 70.2% with Word2Vec. Regarding stock price and Sentiment Analysis correlation, a Classifier with trained data with Logistic regression algorithm can achieve 69.01% accuracy and 71.82% with trained data with LibSVM.

A novel model about the prediction of stock price movement was introduced by utilizing data from Yahoo Finance Message and by using Board SVM with the linear kernel [66]. The researchers' method achieved a better accuracy by 2,08% compared to the model using only historical prices, by 9,83% for stocks that are difficult to predict and by 3,03% compared to human sentiment method.

Finally, the quite new trend of cryptocurrencies was investigated by Steinert and Herff [80] and Matta, et al. [81]. In [80] the authors gathered data from Twitter and by using linear regression they tried to predict the altcoin returns. Depending on the altcoin they made different predictions; their methodology seems to work for explaining very short-term variations in returns for Ethereum, one of the most popular altcoins. In [81] the authors combined 2 million Tweets with Google Trends that was analyzing Bitcoin's popularity. They applied automated Sentiment Analysis (sentistrength) and they found that positive tweets count contributes to the prediction of the movement of Bitcoin's price in a few days.

3.1.2 Product Pricing

A study by Kim et al. [45] investigated whether it is possible to predict food prices by analyzing data from Twitter. Their methodology and the model they provided (Nowcasting model with Regular Expressions and weights) demonstrated a high accuracy in predicting the food prices' fluctuations outperforming Inter-Quartile Range (IQR) filter model, Kernel Density Estimation (KDE) and Auto-Regressive Integrated Moving Average (ARIMA).

Elshendy et al. gathered data from Twitter, Google Trends, Wikipedia, and the Global Data on Events, Language, and Tone database (GDELT), to predict Crude Oil price [57]. Their results showed that the combination of the aforementioned media platforms "*can lead to forecasts for crude oil prices with a reasonably high level of accuracy*". They also identified that these platforms demonstrate different "speeds", as Twitter offers more immediate and fast information (one-day lag), whereas Wikipedia and GDELT are more informative at a two-day lag and Google trends at a three-day lag.

3.1.3 Real Estate

Zamani and Swhartz [10] used a language dataset from Twitter messages (131 million tweets that were mapped to 1, 347 counties), a control dataset of socioeconomic and demographic variables, and an outcome dataset of housing related data. They found a substantial improvement on invaluable real-time indicator for financial markets like prediction on foreclosures and on increased price (13.51% and 18% respectively). Finally, they concluded that data from Twitter is a great addition to prediction models.

Another study utilized data from Twitter in order to understand the shaping and the dynamics of the city of Pittsburgh in the USA [34]. Using clustering, the authors suggested that through local social patterns, like Tweets, it is possible to examine the dynamics of a city (architecture, development, demographics, geographic characteristics, neighborhood and municipality borders, etc.).

3.2 Marketing

Companies and enterprises are trying to capitalize the analysis of data provided by SM [21]. For instance, Vijayaraghavan et al. [44] proposed a method that quantizes community interactions with SM to understand and influence consumer experiences. Kalmer studied the potential of SM analytics in the marketing field and found that due to various biases, SM data cannot be regarded as a consistent and accurate data source [46]. However, when properly implemented "SM Analytics can deliver great value for marketing strategy and business in general".

Gruner et al. examined to what extent SM communication and online advertising are linked to the sales volume of new products [70]. By using Seemingly Unrelated Regressions (SUR) they managed to identify positive, but diminishing relationships, between SM communication and the sales volume and profits of a new product.

In the study by Aiello et al. [90] the friendship prediction and homophily in SM was examined. The authors used data from not very researched SM like flickr, last.fm (an online music database, music recommendation social network) and aNobii (a social network that aims at book readers) and by using several similarity metrics they concluded that "all the user profile features have substantial predictive power" and that users with similar hobbies and interests like music, reading and photography, are more likely to be friends. McGee, Caverlee and Cheng [94] proposed a novel network-based method for location prediction in SM based on tie strength (the level of intensity of a social relationship) by using over 73 million Twitter accounts and 100 million geo-encoded tweets. They identified that there is a correlation between relationships and physical proximity and their approach notably improved the results of SM users' location estimation by reducing the average error distance for 80% of Twitter users by 50%.

Finally, in [95] the authors examined the negative link prediction (the indication of dislike, suspicion or disagreement on opinions) in SM by exploiting positive links (the indication of liking, approval or friendship) and content-centric interactions. Their experimental results on real-world social networks via their proposed framework managed to accurately predict negative links, a solution that can aid SM services such as recommender systems.

3.2.1 Customer needs

Researchers used over 33 million check-ins in Manhattan that were collected from Foursquare through its Twitter API [9]. They studied the concept of ranking, in a model of predicting collective human mobility and urban properties by proposing a methodology that implements Sorensen Similarity Index (SSI), Regression Analysis and Cosine Similarity.

They proposed three (3) different scenarios:

- rank-distance,
- the number of venues in the region, and
- a check-in weighted venue schema to compute the ranks in the model and they achieved 63.84%, 63.94% and 71.35% performance in Sorensen Similarity Index (SSI) and 31.52%, 32.01% and 55.09% R-squared obtained from regression analysis respectively.

According to their methodology, the prediction of human mobility can "open the doors onto novel research about activities within the cities".

A method for predicting future consumer spending from Twitter data was proposed in [16]. The evaluation of the proposed methodology (time series analysis models and machine-learning regression models: SARIMAX, Gradient Boosting Regression, AdaBoost Regression,) demonstrated statistically significant improvements in prediction. The researchers managed to reduce forecast errors from 11% to 18% for a three to seven (3-7) day predicting horizon by using exogenous variables.

He et al. gathered around 476k tweets containing the names of two of the largest retail chains in the world; Costco and Walmart [73]. Using tools like Lexalytics and Leximancer for sentiment analysis via classification and clustering respectively, they examined whether Social Media Analytics can provide potential and insights to companies, in order to improve their competitiveness and solve business issues. According to their results, analyzing just the social media mentions and sentiment trends alone is not enough for linking them with buying trends.

Two location-based social networks (LBSNs) platforms, Gowalla and Foursquare were used in order to predict the next check-in location of individuals [76]. The authors combined data from these two (2) SM with the person's check-in history. According to their results their methodology can predict the next check-in location.

3.2.2 Trends & Entertainment

Ni et al. utilized around thirty (30) million hashtags from Twitter to predict subway passenger flow and detect social events [52]. Their approach, called Optimisation and Prediction with hybrid Loss function that combined Linear Regression with Seasonal Autoregressive Integrated Moving Average (SARIMA), achieved precision 98.27% and recall 87.69% for the baseball games. Based on these data they also proposed an optimization model to forecast subway passenger flow that, compared to existing models, demonstrated an increase in prediction accuracy and robustness.

2.4 million tweets were gathered and analyzed in order to predict Spotify streams for newly released music albums [39]. The author used Linear Regression with Spearman's rank correlation coefficient (Spearman’s RHO) and concluded that the volume of tweets for each album and artist is positively related with Spotify streams and that there is no connection between binary sentiments (positive or negative) with the number of followers and the Spotify stream.

3.2.3 Product Promotion

Hudson et al. used Tweets from France, UK and USA and applied Multiple Regression Analysis with mean-centered brand anthropomorphism and their interaction on Brand Relationship Quality (BRQ) [58]. Their findings demonstrated that there is strong relation between SM and brand relationship quality and that “engaging customers via social media is associated with higher consumer-brand relationships”. According to the study, using SM platforms and data can provide a mechanism for businesses to plan their pricing, marketing and promotion strategies.

Platforms like TripAdvisor, Yelp, Urbanspoon, and Foursquare were used with the aid of Pearson correlation analysis and hierarchical multiple regressions to identify that the number of SM reviews (by customers) has a significant impact in restaurant performance (for instance, the higher is the guests’ overall evaluation, the higher the restaurant sales [48].

Ong and Ito’s work evaluated the effectiveness of SM Influencers marketing campaign that was performed by a Singaporean Tourism Organization [74]. They investigated whether consumer attitude change can be affected and thus predicted by using SM. They concluded that by using SM the marketers can create a “creative interactive and engaging content”.

3.3 Sociopolitical

Many warnings have been raised regarding the role of SM in spreading fake news, false rumors, misinformation, or in extreme cases to be utilized as a propaganda mechanism [35], [37], [59].

Daily, new efforts are being made to leverage SM for prevention of medical and health dangers. In [14] the authors are proposing the creation of a framework for fighting child obesity through SM. In [92] Huh investigated, with the aid of machine learning, unstructured health data of one million Korean citizens using datasets obtained from the National Health Insurance Service and applied text mining, word cloud and web crawling techniques in order to analyze and visualize (providing an application compatible with Android and iOS) keywords from services such as Google Trends, Naver News and About.com for obesity related issues. According to the author's thorough study, big data analysis of obesity healthcare and personalized health activities could provide valuable solutions for deducing obesity.

Many efforts have been made for forecasting cases that could greatly affect citizens. For instance, a study by Wang et al. focused on crime and more specifically, on predicting hit-and-run incidents in Charlottesvile, Virginia, USA [60]. They used a Generalized Linear Regression model (GLR) with Semantic Role Labeling (SRL) and Latent Dirichlet Allocation (LDA), which incorporated data gathered from Twitter within an 8-month period combined with data from criminal incident records, demonstrated a high ability on predicting hit-and-run incidents.

The predictions of health and environmental hazards either caused by people or by the nature itself are of grave importance [13]. McClellan et al. showed that Twitter can be used as a tool to raise public awareness about suicide [47]. Finch et al. focused on Public health implications of SM, during natural and environmental disasters, and they questioned how SM can be used for data prediction and early warnings for disasters [15]. They studied the literature in academic papers based on Facebook, Twitter, Weibo and various environmental disaster terms, like smog, pollution, oil leak, tornado, earthquake etc. One of the conclusions of their study was that "using crowdsourcing from social media sites provided a lot of benefits to emergency responders and enhanced public health knowledge and surveillance". Other researchers surveyed data analysis and management in disaster situations organizing under specific Computer Science disciplines like: data integration and ingestion, information extraction, information retrieval, information filtering, data mining and decision support [84].

Our study mainly focus on cases were the results can affect millions, like elections, diseases and natural phenomena.

3.3.1 Elections

In [49] the authors created a model based on Thelwall's SentiStrength emotion detection algorithm [62] and by using around 14 million Tweets they tried to predict the outcome of the UK general elections. They projected a parliament with 285

Conservative party seats and 306 Labour Party seats. Their prediction proved inaccurate as the Conservatives won the election with 330 seats compared to 232 by the Labour party.

The authors of [23] proposed *"a computational public opinion mining approach"*, they gathered data from Twitter and they implemented machine learning classifiers, Linguistic Inquiry and Word Count (LIWC) and they applied the Mallet implementation of latent Dirichelt allocation (LDA) in order to explore detect and analyze economic issues in SM during the 2012 US Presidential Elections. Their methodology, calculating the difference between the number of positive topics and the number of negative topics (DPNT), indicated that Obama had the advantage over Romney, and it validated and even outperformed the Pew Research Center survey results [63].

In [50] the authors used thirteen (13) different variables that were available online including Tweets, Celebrity Tweets and Celebrity Sentiments, Twitter Followers, Facebook Page Likes and Wikipedia Traffic for the 2016 US Presidential Elections. They used Linear Regression Analysis with data partitioning and scoring. According to their results, Wikipedia page traffic for the Democratic Party candidates did not correlate with the outcome of the DP election, since Sanders lost to Clinton even though his Wikipedia page traffic was at least two times larger than Clinton's for a period of 2.5 months from November 2015 up to mid-January 2016. The research found correlations between polls and Facebook page likes, and between polls and Twitter. Finally, they concluded that "Machine learning models with linear regression can produce predictions with meaningful accuracy".

In [51] a research about whether SM can predict election results in New Zealand demonstrated that the number of Facebook friends and Twitter followers as explanatory variables, are not good indicators as only 16.7 and 5.4 percent of election winners were predicted correctly respectively. The researchers specified two (2) regression models: i) a linear ordinary least squares (OLS) model of vote share; and ii) a logistic regression model in order to test the validity of their proposed methodology.

Finally, MacDonald and Mao [43] not only used data from SM but also from other resources, like Google trends, Wikipedia, Polls, and news. Text mining techniques in conjunction with vector autoregressive (VAR) methodology were combined into a framework to predict the results of the 2015 Scottish and UK elections with quite impressive results as their prediction was very precise. They predicted correctly the rank of the parties and they have had in many occasions very accurate predictions within decimals (i.e. the mean rate for the percentage of the Conservative party in Scotland was 14.73% and the actual one was 14.90%).

3.3.2 Diseases – Health – Vaccines

Chen et al. [13] dealt with forecasting smog-related health hazard by using SM data from weibo.com. Their proposed Artificial Neural Network, using Extreme Learning Machine (ELM) and Back Propagation (BP), for forecasting the health hazard performed better when it combined both Physical Sensor Data (from ground and satellite sensors) and SM data.

In [33] the authors used data from SM in order to characterize dietary choices, nutrition and language in areas with poor diet and diet-related health issues, also known as "Food Deserts". Their regression model provided a high accuracy (>80%) and improvements over baseline methods by 6-14% in Improving Identification and Surveillance of Food Deserts.

In [47] the researchers proposed a model that is valid for identifying periods of heightened activity on Twitter related to behavioral health using ARIMA. The authors collected 176 million tweets from a 4-year period (2011 to 2014) related to suicide or depression. They managed to identify spikes in tweet volume following a behavioral health event often lasting for less than 2 days. Their model can be used by health organizations to identify periods of heightened mental health–related activity on Twitter and take on prevention and treatment initiatives.

Sadilek et al. investigated whether the prevention of Foodborne Illness is possible by applying data mining on SM [53]. They developed a tool called nEmesis that uses data from Twitter for the Las Vegas area, and they used an SVM classifier. They collected around 16,000 geo-tagged food venue related tweets, along with subsequent tweets from the same users for a five (5) days period. Their methodology was evaluated to be 63% more effective at identifying problematic venues than the current state of the art and it can prevent over 9,126 cases of foodborne illness and 557 hospitalizations in Las Vegas annually.

Santillana et al. proposed a methodology by combining data from four (4) different sources: i) Twitter, ii) Google searches, iii) nearly real-time hospital visit records, and iv) a participatory surveillance system and applying a Linear autoregression exogenous (ARX) model [67]. The researchers managed to demonstrate that the combination of these four (4) sources produces better results by using each source independently. The researches claim to have accomplished predictions one (1) week ahead of Google's Flu Trends real-time estimates, with comparable accuracy, and accuracy of two (2) to three (3) weeks forecasting estimates.

Subramani et al. used text mining and real time analytics on data retrieved from Twitter, to which they applied automatic classification with logistic regression models for predicting Hay Fever in Australia [69]. According to their results, predicting Hay Fever outbreaks is plausible as there is positive correlation between Evaporation, Relative Humidity, Average Wind Speed and Hay Fever tweeting.

The authors of [71] investigated whether it is possible to detect binge drinking in students by using data retrieved from Facebook. The data were not only in textual format, but also included images and video. The study concluded that by using classification techniques it is feasible to create prediction models.

In [72] the authors gathered 10,000 tweets and by applying the Probabilistic Latent Semantic Analysis (PLSA) classification model they tried to predict the box office Performance of 14 Bollywood movies. Their model had a good performance as their prediction was approximately the same as the actual results.

Zhan et al. in concentrated on finding e-cigarette usage patterns [75]. They collected around 333k related posts from reddit and provided users with questionnaires. The study examined many aspects of e-cigarette usage; one of them was the identification of intentions of using e-cigarette flavours and liquids, with the aid of regression. According to their findings, the results of product evaluation can be enhanced by combining data from SM and the questionnaires and the "e-liquid rating increase was associated with an increase of e-cigarette future use intention odds ratio".

Su et al. used a regression model for tweet specificity prediction [77]. They applied their model for mental analysis purposes, and they managed to successfully identify people with moderate or severe depression.

A research also related to detecting depression via user-generated data combined with SM data and most specifically, Instagram, was conducted by Ricard et al. [78]. 749 participants were selected, questionnaires were handed out and access to their Instagram accounts was allowed. The results showed that this data combination can be informative for depression prediction among SM users.

In [79] the authors proposed a methodology for identifying dangerous video challenges by processing Tweets that contain videos. After testing their model, it returned accuracy of 87% in distinguishing dangerous and non-dangerous videos, thus promoting public safety. Other researchers proposed and analyzed new methods for identifying fire in videos using various experiments illustrating the applicability of their method [85].

Vaccines

In this study we also research the role of SM on the spread of the anti-vaccination movement. A major research has been conducted by Tomeny et al. regarding the temporal trends, the geographic distribution and the demographic connection of anti-vaccine beliefs in Twitter, by amassing around 550,000 tweets, for a 6-year period from 2009 up to 2015 [35]. By implementing Least Absolute Shrinkage and

Selection Operator (LASSO) regression, they demonstrated that the volume of anti-vaccine tweets has been steady from 2009 to 2014 and there were noticeable spikes caused by vaccine-related news, such a measles outbreak in California that commenced in December 2014. However, the last spike of anti-vaccine tweets, followed the outbreak. Therefore, the research did not show a connection between the tweets and the outbreak and failed to predict it.

A study in Italy tried to explore the connection between MMR vaccination coverage and on-line search trends and SM activity for a 5-year period from 2010 and 2015 [36]. They implemented Classification and Bivariate Pearson's Correlation Analysis. The results showed a significant negative correlation between the MMR vaccination coverage and Internet search query data (p = 0.043), tweets (p = 0.013) and posts (p = 0.004). Therefore, as the authors concluded, it was not possible to demonstrate if the decrease of the percentage of Italians that are been vaccinated. What was observed after 2013 is related to the increase of the anti-vaccination tweets in Twitter and posts in Facebook. However, it was established that the triggering events for the outbreaks were two (2) court decisions in the country that linked autism with vaccination: i) on April 2012, the Italian Court of Rimini ruled that there is a link between the MMR vaccine and autism and ii) on November 2014, an Italian court in Milan awarded compensation to a boy for vaccine-induced autism. However, the authors did not incorporate the two (2) incidents into their methodology.

In [37] the authors collected data from Twitter for a period of one (1) month before and two (2) months after a measles outbreak in Disneyland. With the aid of supervised machine learning classifiers: relevance, sentiment bearing and sentiment polarity, the authors managed to identify that the number of vaccine-related tweets messages during the outbreak increased, with positive messages "dramatically increasing". However, according to the graph that they provided for this 3-month period, the negative vaccine tweets are, constantly, at least twice in volume as the positive ones, and there is no observation of an increase to vaccine-related tweets prior to the outbreak.

Radzikowski et al. presented a quantitative study of Twitter narrative after a 2015 measles outbreak in the USA [38]. They collected around 670,000 tweets from across the globe, referring to vaccinations from the 1st of February 2015 and for a 40-day period. They identified the dominant terms, the communication patterns for retweeting, the narrative structure of the tweets, the age distribution of those involved and the geographical patterns of participation in the vaccination debate in social media. The most important result from this research was that there is a strong connection between the engagement of Twitter users in vaccination debates and non-medical exemption from school-entry vaccines. More specifically they provided evidence that "*Vermont and Oregon with the highest rates of exemption from mandatory* *child school-entry vaccines had notably higher rates of engagement in the vaccination discourse on Twitter*". However, these tweets were collected after the measles outbreak and therefore, there is no evidence that the outbreak could have been predicted.

Table 2: Summarizing Social Media Prediction attempts

Citation	*Category*	*SM Platforms*	*Number of records*	*Pre-processing / cleaning*	*Algorithms - Methodologies*	*Valid Predictions*	*Key Findings*
Finance – Stock Market							
[4]	Enterprise Credit Risk	Hexun.com, Finance.sina.com.cn	11,034 tweets	Minimum 30 posts limit, financial data anomalies removed	Regression	79.13%	Opinions extracted from SM (posts and commentaries) surpass the opinions of analysts regarding risk prediction
[6]	Fraud Prediction	Seeking Alpha, Twitter	SM content from 192 firms	Numerical digit and special symbols dropped, all words converted into lower case, stop words removed	Document Classification, SVM with Gaussian Kernel	83,57%	Combining SM features with a set of financial ratios makes the fraud prediction accuracy at 83.57% on average
[7]	Trading	StockTwits	45 million messages	R library tm used, data separeted according to AM or PM post time, non-words and stop-words removed, words converted into lower case	Sparse Matrix Factorization	51.37%	Market-timing predictions using the SMF model and StockTwits streams perform better than most basic baseline models
[11]	Financial Markets	Twitter	46,444 tweets	no mention, SentiStrength used	Sentiment Classification (SentiStrength)	Partially	Negative correlation of stock market prices and positive trading volume with mood indicators and tweet volume
[17]	Financial Markets	Google search volumes, Bloomberg news and Twitter	401,923 Bloomberg articles, no info about GT, Twitter	lemmatization, sentence boundary detection, part of speech tagging and noun phrase detection, custom country level dictionary	Delta Naive Bayes (DNB)	F1 measure can be as high as 0.407	Multi-source predictions consistently outperform single-source predictions. "Twitter bursts, if properly processed can be an invaluable real-time indicator for financial Markets"
[22]	Financial Markets	Twitter	250,000 tweets	Tokenization, Stopwords removal and regex matching for removing special characters	Sentiment Analysis (SA) with Classifier using N-gram and Word2Vec representations – Classifier with Logistic regression (LR) & LibSVM	Accuracy: 1) 69.01-71.82%	"A strong correlation exists between rise/fall in stock prices of a company to the public opinions or emotions about that company expressed on twitter through tweets"

[66]	Financial Markets	Yahoo Finance Message Board	no specific number of records mentioned	stop words removed, all the words lemmatized by the Stanford CoreNLP	SVM with the linear kernel	More than 60% accuracy	The method predicts if the price of the stocks is increasing or decreasing.
[80]	Cryptocurrency	Twitter – Prices	426,520 tweets	included altcoins that were referred to on Twitter on at least 10% of all days	Least squares linear regression	Partially	There is a connection between SM activity and sentiment on Twitter and altcoins
[81]	Cryptocurrency	Twitter – Google Trends	1,924,891 tweets	Comparing the timeline of tweets and the fluctuations in the Bitcoin market, the specific day that provide a better correlation value was determined	Automated Sentiment Analysis	Yes	positive tweets-count contributes to the prediction of the movement of Bitcoin's price in a few days.
Finance – Product Pricing							
[45]	Food	Twitter	78,518 tweets	ambiguity (words with more than one meaning) and redundant messages or spam bots removal	Nowcasting model with Regular Expressions & weights	Yes, outperforming ARIMA, ICQ and KDE	"the volume of tweets mentioning food prices tends to increase on days when food prices change sharply".
[57]	Oil	Twitter, Google Trends, Wikipedia, GDELT	no specific number of records mentioned	Missing values during weekends and public holidays excluded & missing data during working days interpolated, in order to create a 5 days per week observation time. Lags were limited to a maximum of 3 days	ARIMA and ARIMAX time series forecasting models	Partially	Combination of the many SM platforms can provide a more accurate Crude Oil price prediction.
Finance – Real Estate							
[10]	Foreclosure and prices	Twitter	131 million tweets	tokenize, stem, and remove punctuation and standard stopwords from the associated utterances	Combination of control model and language model	Improvement of foreclosures (from r = .37 to r = .42) & increased price (from r = .50 to r = .59)	Twitter is adding predictive information about the Real Estate market compared to traditional socioeconomic predictors.
Citation	*Category*	*SM Platforms*			*Algorithms - Methodologies*	*Valid Predictions*	*Key Findings*
Marketing – Customer Needs							

[9]	Human Mobility	Twitter	33 million check-ins	employed the number of venues located in a circle centered at the origin with a radius equal to the distance between the origin and destination to compute the ranks. Use check-in data for weighting the venues	Sorensen Similarity Index (SSI), Regression Analysis, Cosine Similarity	55.09%-71.35%	SM check-ins play a significant role in improving the predictability of mobility patterns
[16]	Consumer Spending	Twitter	68,730 messages	The text of each tweet is cleaned (any material outside of the grammatical text is removed) and processed with a part-of-speech tagger. PoS tag patterns are then applied to extract the head noun	ML regression (SARIMAX, AdaBoost Regression, Gradient Boosting Regression)	11%-18% Error Reduction	Combining SM data as exogenous variables alongside lagged values of the consumer spending index improve performance
[73]	Customers' Product Sentiments	Twitter	475,959 tweets	keyword querying	Classification - Clustering	Plausible	Analysing only social media mentions and sentiment trend are not sufficient
[76]	Next location check-in	Gowalla - Foursquare	118,418 users, 1,463,937 locations, 7,828,113 check-ins, 997,401 social links	select users who have at least 80 checkins and remove POIs that have fewer than 20 check-ins	k-nearest neighbors' algorithm	Yes	The methodology performs well in predicting the next check-in location
Marketing – Trends & Entertainment							
[52]	Events	Twitter	29.7 million geo-tagged posts	hashtag-based method to remove unrelated tweets, Remove stop words, punctuation and duplicated strings	Linear Regression with SARIMA	Precision 98.27%	Baseball games were predicted by using subway passenger flow
[39]	Spotify Streams	Twitter	2,469,574 tweets from 898,837 unique users and 5,804,407 spotify streams	keywords from albums not released on Spotify removed, extreme outlier removed, Replace all capitalization with lower case. hashtags, @mentions, URLs and stop words (top 25) removed, Album title/artist name filtered out	Linear Regression with Spearman's rho	Partially	There is a strong relationship between the volume-related Twitter variables and Spotify streams.

[72]	Movies Performance	Twitter	10,269 tweets and retweets	no mention	Probabilistic Latent Semantic Analysis (PLSA) classification model	Yes	the predicted values were approximately the same as the actual values
Marketing – Product Promotion							
[58]	Consumer and brand relation-ships	Twitter	236, 207 and 281 users for 3 studies	no mention	Multiple Regression Analysis with mean-centered brand anthropomorphism and their interaction on BRQ	Partially	SM interaction has a positive association with BRQ and this relationship depends on brand anthropomorphism
[48]	Restaurant Performance	TripAdvisor, Yelp, Urbanspoon & Foursquare	7,935 reviews	no mention, 5-point likert-scale used	Pearson correlation analysis and hierarchical multiple regressions	Plausible	SM reviews by customers has a significant impact in restaurant performance
[74]	Tourism	Twitter – Facebook - Questionnaire	200 users	no mention, 5-point likert-scale used	Regression	Yes	SMI can cause an attitude change in the consumers and can affect their behavioural intention
Sociopolitical – Elections							
[49]	Elections	Twitter	13,899,073 tweets	Where a tweet contained more than one of the search terms, the tweet from the sample to avoid misallocating was removed	SentiStrength emotion detection algorithm (Classification) [62]	No	limitations of using Twitter to forecast in multi-party systems where there is a 'majority' regionalist party
[23]	Elections	Twitter	24 million tweets	remove stopwords	ML classifiers, Linguistic Inquiry and Word Count (LIWC) applied the Mallet implementation of Latent Dirichlet Allocation (LDA)	Yes	The difference between the number of positive topics and the number of negative topics indicated that Obama was having the advantage over Romney
[50]	Elections	Twitter, Facebook, Wikipedia	no specific number of records mentioned	no mention	Linear regression analysis with data partitioning and scoring	10 % lower root mean squared error	ML models with linear regression can produce predictions with meaningful accuracy

[51]	Elections	Twitter, Facebook	Election Day observations of friend/follower counts (161 Facebook, 98 Twitter)	one outlier in the Twitter dataset was removed. "fan pages" which prevented people from "friending" opponents removed	2 regression models were specified: (1) a linear OLS model, (2) a logistic regression model	No	Increases in SM following may be related to increased numbers of votes, but not in sufficient quantity to change election results as most voters are not SM users.
Sociopolitical – Health/Diseases							
[13]	Smog	Weibo and Physical Sensor Data	315 million tweets with their retweet and like records	scaling up the fraction of rare tweets, social network diffusion	Extreme Learning Machine (ELM) and Back Propagation (BP)	Better Performance	Combination both social media data and physical sensor features improves prediction of high-level smog
[33]	Diet related	Instagram	4 million posts from over 8 million users	Posts with no matches to US Department of Agriculture descriptors were disregarded. filtered tags with valid latitude-longitude information LDA was used	Regression	Accuracy >80%	High accuracy (>80%) and improves over baseline methods by 6-14%
[47]	Mental Health	Twitter	176 million tweets	similar terms were added to the query as exclusion criteria, anomaly in the series was removed	AMIRA	Partially	Spikes in tweet volume following a behavioral health event often last for less than 2 days
[53]	Health	Twitter	15,900 geo-tagged tweets from 3,600 users per day over 3 months	cleanup, filtering, snapping and labeling	SVM Classifier	64% improvement	Improvement compared to public health controls. Can prevent over 9,126 cases of foodborne illness and 557 hospitalizations in Las Vegas annually.
[67]	Health	Google searches, Twitter, hospital visit records, and participatory surveillance system	12,000 tweets annotated for relevance (no other mention of total size of data)	distinguished tweets that indicated an infection rather than discussing influenza in other contexts	Linear autoregression exogenous (ARX) model	Yes	Combining multiple sources produces better results by using each source independently. Predictions one week ahead of Google's Flu Trends were made.
[69]	Health – Hay Fever	Twitter	681 tweets	the standard pre-processing tasks were implemented	Automatic classification with logistic regression model	Partially	The impact of specific variables that assist future forecasting were identified

[71]	Health – Binge Drinking	Facebook	4,266 posts	two separate rounds of annotations to determine whether or not each post contained related contents. Emoticons, links, hash tags, and special characters filtered out. The link was replaced with the URL. Stop words were also filtered out using the stop word list from the Natural Language Toolkit	SVM classifiers with linear kernel	Feasible	Detecting drinking-related posts on SM is plausible using a combination of text and image/video classification techniques
[75]	Health – e-cigarette	Reddit - Questionnaire	332,906 posts	Levenshtein distance to identify misspelling brand names was used. Words with Levenshtein distance less than or equal to 2 were manually checked with the help of Google search. In the analysis only, keywords mentioned at least 3 times from the survey answers were included	Regression	Yes	e-liquid rating increase was associated with an increase of e-cigarette future use intention odds ratio
[77]	Language specificity - Depression	Twitter	7,330 tweets sampled from 3,665 users	Re-tweets are excluded, URLs and usernames are replaced by special tokens, all emojis are preserved	Regression	Yes	Applying the language specificity prediction can identify people with depression
[78]	Health – Depression	Instagram – Questionnaire	749 users with 333.55 mean number of posts	a bag-of-words model was used, English stop words were removed using the Natural Language Tool Kit library, processed captions and comments of a user's most recent 20 posts were aggregated and converted to a feature vector	linear regression model with an elastic-net regularization	Yes	Combining SM and user-generated data can be informative for depression prediction among SM users.
[79]	Public Safety	Twitter	25k tweets	scraped the tweets and selected those containing at least one of 10 specific hashtags	Cohen's kappa classification	Yes – 87% accuracy	Able to identify potentially dangerous challenge videos, and promote public safety
Sociopolitical – Vaccines							

[35]	Vaccines	Twitter	549,972 tweets	tweets were geolocated, retweets were geocoded to the area from which they were retweeted, wording was changed slightly to maintain user anonymity	LASSO regression	No	Heavy SM traffic after the outbreak.
[36]	Vaccines	Twitter, Facebook, Google Search Queries	no specific number of records mentioned	specific word removed	Classification and Bivariate Pearson's Correlation Analysis	No	Heavy SM traffic after the outbreaks. There were 2 major incidents (court decisions) that triggered the outbreak. Heavy SM traffic after the court decision.
[37]	Vaccines	Twitter	no specific number of records mentioned	tweets were normalized using the maximum number of relevant tweets in a 140-day window, Tweets were labeled for relevance to the topic	Three supervised machine learning classifiers: relevance, sentiment bearing and sentiment polarity	No	Increase in the number of vaccine-related tweets messages during the outbreak, with positive messages "dramatically increasing".
[38]	Vaccines	Twitter	669,136 tweets	stop words removed, specific words excluded, words with hashtags were considered as different words	Association Rules	No	Heavy SM traffic after the outbreaks. States in the USA with the highest rates of exemption from mandatory child school vaccines had notably higher rates of engagement in the vaccination discourse on Twitter.
Sociopolitical – Natural Phenomena							
[42]	Earthquakes and Disaster measurement	Twitter	52.55 million messages from 13.75 million unique users	each keyword is considered separately, aggregated the tweets by location and used time stamps for temporal analysis	Classification and total absolute score, word-count normalized score, and trinary classification.	Partially	An efficient tool for measuring the physical damage caused by a disaster.
[82]	Extreme Weather	Twitter – Weather Forecasts	280k tweets	content in each monitored stream was processed through a Language Analysis pipeline, fast text tool	Classification	Partially	social media data can be used together with probabilistic weather forecasts to automatically detect an ongoing event

3.3.3 Natural phenomena

Despite that there are some prediction tools that alert the public about a forthcoming natural disastrous phenomenon using data from Social media, like Toretter that was created in Japan and provides a faster notification than the Japanese government's Meteorological Agency, not much success has been accomplished on the specific field [55],

[56]. Goswani et al. have surveyed the importance of using data mining in order to combat natural disasters and they have produced a literature study with many relevant papers dating before 2015 [68]. Finally, in [65] the authors reviewed how data mining can aid to predict, detect and develop disaster management strategies by collecting data from Twitter. They also proposed a system for creating a "disaster management database" where by incorporating data from i) Twitter, ii) news and iii) other SM and internet sources, information can be gathered and processed so that natural disasters can be effectively predicted.

Whilst SM are great as information tools, they can add unnecessary anxiety and information ambiguity as a study has shown [59]. In this study, the rumor dynamics by analyzing Tweets regarding the 2010 Haiti Earthquake was explored and the key finding was that the anxiety and confusion SM can create on extreme cases like earthquakes, can be minimized by linking the tweets with "websites of the emergency response center or authenticating governmental organizations, RSS, streaming videos, photo, text message, Retweets, etc. "

In [42] the authors demonstrated the usefulness of SM both in assessing the disaster damage and at the same time as a predictor for the damage inflicted in the areas affected by natural disasters. They implemented Classification on data retrieved from Twitter with total absolute score, word-count normalized score (relative score), and trinary classification (+1, −1, and 0). The study also proved that online SM activity is related to the proximity of the region of the path of the hurricane. Their methodology demonstrated that online response reflects better the damage and the path of the hurricane than the information from Federal Emergency Management Agency (FEMA) which is responsible for coordinating the response to a disaster in the USA.

Sakaki et al. provided a mechanism for detecting Earthquake activity from Tweets [55]. Their proposed algorithm (Classification with SVM) demonstrated remarkable results with high 'Location estimation accuracy of earthquakes' and 'Trajectory estimation accuracy' of a typhoon solely from tweets. They created an alert system via email, and they managed to send alarm notifications within twenty (20) seconds to one (1) minute and their delivery time was faster than the broadcasts of the Japan Meteorological Agency (JMA).

Researchers gathered data from Weather Forecasts Social Media Monitoring, Event Detection on Social Media Streams and Informativeness Classification of Social Media Content [82]. According to their results "social media data can be used together with probabilistic weather forecasts to automatically detect an ongoing event".

4. Statistical Analysis

In this paper we examined and analyzed 43 research papers with the following characteristics:

1. They are recent.
2. They used data from SM.
3. They relate to three (3) fields (Finance, Marketing, Sociopolitical).

4. They attempted to predict outcomes.

4.1 General Statistics

All the papers are dated from the past four (4) years. The distribution is depicted in *Table 5*.

Table 3: Number of Sources and Types of Social Media Used

		Number of Sources			**Social Media**						
Category	Analyzed	One	Two	More than two	Only Twitter	Other without Twitter	Twitter and other	Facebook	Google trends and search	Wikipedia	Instagram
Finance	**12**										
Stock Market	9	5	3	1	3	3	3		2		
Product Pricing	2	1		1	1		1		1	1	
Real Estate	1	1			1						
Marketing	**10**										
Customer Needs	4	3	1		3	1	1				
Trends	3	3			3						
Promotion	3	1		2	1	1	1	1			
Sociopolitical	**21**										
Elections	4	2	1	1	2		2	2		1	
Health	11	9	1	1	5	45	1	1	1		2
Vaccines	4	3		1	3		1	1	1		
Natural Disasters	2	1	1		1		1				
Total:	43	29	7	7	23	10	10	5	5	2	2

Table 4: Algorithms Used and Prediction Achieved

		Algorithms				**Is Prediction achieved?**		
Category	Analyzed	Classification	Regression	Combination Classification & Regression	Other (Association/ Clustering) and combinations with Regression or Classification	Yes	No	Partially/ Plausible
Finance	**12**							
Stock Market	9	4	2	2	3	6		3
Product Pricing	2		2			1		1
Real Estate	1		1			1		
Marketing	**10**							
Customer Needs	4	1	2	1	1	3		1
Trends	3	1	2			2		1
Promotion	3		3			1		2
Sociopolitical	**21**							
Elections	4	2	2			2	2	
Health	11	4	5	2	2	8		3
Vaccines	4	2	1	1	1		4	

	Natural Disasters	2	2						2
Total:		**43**	16	20	2	7	24	6	13

Table 5: 4-year distribution

Year	**2015**	**2016**	**2017**	**2018**	**2019**
Papers per year	5	15	12	9	2
Percentage	11.63	34.88	27.91	20.93	4.65

Twenty-nine (29) out of the forty-three (43) papers (67.44%) used only one (1) source of data for their predictions, whilst 16.28% used ten (10) different sources and the remaining 16.28% incorporated data from more than two (2) different SM platforms. The main data language used was presumed to be English in thirty-eight (38) papers, whilst in two (2) the authors conducted their research in data written in Chinese [4] and [28], two (2) in Italian [36] and [82] and in one occasion [45] in Bahasa (Indonesian language).

Regarding the pre-processing and cleaning of data, our analysis showed that methods like non-words and stop-words, punctuation and duplicated strings, ambiguity (words with more than one meaning) and redundant messages or spam bots removal, regex matching for removing special characters, words converted into lower case were implemented. In addition, thresholds and limits in dates, occurrences and distance were imposed. There were six (6) studies were there was no mention of any cleaning or pre-processing methods.

4.2 Social Media Used

As depicted in Table 3, Twitter was the prominent SM platform of choice for the studies, with the number of the processed tweets range from a few hundred (681) up to 315 million. With a dominating 76.74% the researchers used data from Twitter's API, either by itself (53.49%) or combined with other SMs (23.26%). Data from Facebook were employed in five (5) occasions, in five (5) from Google search queries and trends, twice from Wikipedia and Instagram. In two (2) occasions, the study was conducted by using the most popular Asian SM.

Tweet volume

There is a very wide spread for the number of tweets that has been processed by the forty-three (43) studies in order to make SM predictions (Table 6). The minimum number was 681, whilst the maximum was around 315 million. Most studies collected 10k to 100k tweets (31.03%) and more than 10 million (34.48%).

4.3 Methodology and Algorithms Used

Regression algorithms seems to be the most popular choice by the researchers as twenty (20) out of the forty-three (43) papers used them (46.5%). In 37.2% of the studies Classification algorithms were used, whilst in two (2) occasions a combination of Classification and Regression was implemented. For the remaining seven (7) papers (16.3%), Association rules or Clustering or combination of other algorithms with Regression or Classification was used.

4.4 Observations

Our data analysis can be summarized in a number of observations:

Observation 1: We have six (6) cases of no prediction. Two (2) of them are elections and four (4) of them are vaccines. All of them belong to the socio-political field and the source of data is Twitter or Twit_plus.

Observation 2: There are thirteen (13) cases of partial prediction. Four (4) belong to the field of finance, four (4) to the field of marketing and five (5) belong to socio-political. All of them have Twitter or Twit_plus as source of data, and only one belongs to the category of Health and has Facebook as source of data.

Observation 3: In the category of vaccines we have no prediction at all, summing up to four (4) cases.

Observation 4: In the field of socio-political with data from Twitter or Twit_plus there is no prediction at all, summing up to six (6) cases.

Observation 5: When the source of data is Instagram, we find out that there are two (2) cases of prediction.

The number of our observations help us pinpoint, evaluate and discuss the actual answer (based on data observation) that is stated in 4.5.

4.5 Prediction

The specific question, whether a correct prediction was achieved is quite difficult to answer, as there were occasions were even a slight success was considered as a prediction. At the same time, additional evaluations with different datasets must be carried out in order to consolidate the methodology as *accurate* and *replicable*.

However, according to the authors of the examined studies, more than half of the methodologies (53.1%) managed to achieve a valid prediction.18.8% of the proposed methodologies failed to achieve an accurate prediction, whereas, on the remaining 28.1% a useable prediction was characterized as plausible or at least managed to be validated partially.

Prediction success according to number of Twitter posts

There is no connection between achieving a correct prediction and the number of tweets processed by the studies as shown at Table 6. Leaving aside the volume of less than 10k and 1 to 10 million tweets, as the sample is very small (just 2 occurrences), regardless if the number of tweets collected by the researchers are between 10k and 100k or more than 10 million, the correct prediction is almost the same, 66.67% and 60% respectively.

Table 6: Processed number of tweets and prediction success

Number of tweets	Number of articles	Yes	No	Plausible
less than 10k	2 (6.90%)	1 (50.00%)		1 (50.00%)
10k - 100k	9 (31.03%)	6 (66.67%)	2 (22.22%)	1 (11.11%)

100k - 1 million	6 (20.69%)	3 (50.00%)		3 (50.00%)
1 million to 10 million	2 (6.90%)	2 (100.00%)		
More than 10 million	10 (34.48%)	6 (60.00%)	1 (10.00%)	3 (30.00%)
Total:	29	18	3	8

Prediction success according to number of sources

As it is depicted in Table 7, when two (2) sources are been used in the methodology, the prediction was accurate in 71.43% of the occasions. This is a quite interesting find, as there are studies that concluded that combination of many SM platforms and exogenous data can provide a more accurate result [13], [57], [67], [73], [78], [82]. However, when using more than two (2) sources, the prediction was accurate in 42.86% of the studies.

Table 7: Used number of sources and prediction success

Number of sources	**Number of articles**	**Yes**	**No**	**Plausible**
1	29 (67.44%)	16 (55.17%)	4 (13.79%)	9 (31.03%)
2	7 (16.28%)	*5 (71.43%)*	*1 (14.29%)*	1 (14.29%)
more than 2	7 (16.28%)	3 (42.86%)	*1 (14.29%)*	*3 (42.86%)*
Total:	43	24	6	11

Prediction success according to Twitter

Table 8: Twitter as source and prediction success

Twitter or not	**Number of articles**	**Yes**	**No**	**Plausible**
Twitter	23 (53.49%)	11 (47.83%)	4 (17.39%)	*8 (33.78%)*
Twitter and other	10 (23.26%)	5 (50.00%)	*2 (20.00%)*	3 (30.00%)
No Twitter	10 (23.68%)	*8 (80.00%)*		2 (20.00%)
Total:	43	24	6	11

As it is shown in Table 8, the best performance on having accurate prediction (yes) was found when Twitter was not involved (80%). That's another intrigued find as Twitter is the most popular SM platform used for SMP (in approximately 77% of the 43 studies). This assumption echoes and validates the issues occurred when researchers were using Twitter data and that were identified by the majority of our reviewed studies.

Prediction success according to algorithm

Combining regression, classification and other methods seem to work better in 71.43% of the studies (Table 9), while classification seems to underperform with less than half accurate predictions (47.83%)

Table 9: Algorithms used and prediction success

Algorithms	Number of articles	Yes	No	Plausible
Classification	16 (53.49%)	7 (43.75%)	*3 (18.75%)*	*6 (37.50%)*
Regression	20 (23.26%)	12 (60.00%)	2 (10.00%)	6 (30.00%)
Other and combination	7 (23.68%)	*5 (71.43.00%)*	1 (14.29%)	1 (14.29%)
Total:	43	24	6	11

5. Discussion

SM is flourishing due to the constantly increasing number of users and their user generated data. However, despite the effort of educating and training user from elementary students, even elders [93] for the proper use of computers and SN, like for instance the development of a novel and innovative competency-oriented Social Multimedia Computer Network Curriculum by Huh and Seo [91], there are many studies that demonstrate misuse [103], [104], [105], [107] of SM.

In this paper we surveyed and categorized the latest attempts on Prediction using SM. In [12] the authors mentioned the limitations of predictions through SM data and they checked the hypothesis if SM can be used for predictions. They concluded that "predictions can indeed be made using SM data, however with caution due to "the limitations which were outlined".
We carried out a literature review on three (3), of the many, fields that SM prediction can be utilized, along with some significant categories and examined whether prediction analysis works on these fields:

- *Finance* (Stock Markets, Product Pricing, Real Estate),
- *Marketing* (Customer Needs, Trends & Entertainment, Product Promotion) and
- *Sociopolitical* (Elections, Health/Diseases/ Vaccines, Natural Phenomena).

Table 2 summarizes the most representative SMP attempts we discussed in this publication.

According to the studies we surveyed there are significant issues and drawbacks with SM prediction. However, in some fields the research demonstrated that it is easier to predict. More specifically when numbers are involved prediction is feasible. Especially in the Finance the studies demonstrated that the prediction of stock market, product and real estate prices can be enhanced by using data from SM.

Mixed results were identified at the marketing field. The results from the studies were split as there were indications that a prediction could be accomplished in some cases whereas in some of those studies the results were inconclusive.

The most troublesome field was the effects on "Sociopolitical". At the election fields there were some grand successes, like [43] and some memorable disastrous predictions like [49].

Even more strenuous proved to be the categories where disasters from natural phenomena and health hazards were to be predicted. Even though there were studies that aided to the enhancement of hazardous public health issues like smog [13] and food poisoning [53], predicting natural phenomena disasters and virus outbreaks was nearly impossible.

More specifically, regarding the antivaccination movement issue, the review of the literature demonstrated that there is an increase to the number of vaccine related terms in posts at SM after a disease outbreak. What was evident though, was the correlation of real-life situations and virus outburst. Court decisions in Italy that adjudicated compensation because of MMR and a Chinese vaccine distribution scandal resulted in the reduction of vaccination in kids therefore, to an increase in virus-related symptoms.

Our survey demonstrated that the vast majority, more than 75%, of researchers use data from Twitter, and more than half prefer to use Regression algorithms for the predictive model.

Finally, all the authors of the studies under examination proved via the evaluation of their studies that combining data from more than one source, like data from more than one (1) SM platforms and/or from Google search queries and trends and News, improved the prediction results. Our survey echoed their result regarding the combination of two (2) different sources, whilst using more than two wasn't too effective.

6. Conclusions - Future Work

Whilst data from Twitter is a very popular choice for prediction, there are many times that they simply do not work, as Gayo-Avello has profoundly indicated [64]. Therefore, other SM should be considered in order to achieve a better prediction performance. It should be noted that social influence can be incorporated for Entertainment &Trends prediction, like in the music industry [86]. However, Twitter data are more readily available via an API, and recent studies indicate its potential, particularly in the election prediction arena [83].

Regarding the disease outbreak prediction, there is still no evidence that increased SM traffic can lead to an outbreak and there is no forecasting model that could predict the outbreak based on SM, so far. Therefore, we are currently working on a framework that could:

A. Collect live spatio-temporal data regarding public opinion on vaccinations and their adoption and form a *monitoring system*.

B. Analyze and *classify SM users' weighted opinions*.

C. *Classify SM users* based on their beliefs on vaccinations.

D. Conduct *sentiment analysis* (positive, negative, neutral).

E. *Intervene* in support of efficient and effective disease prevention and control, using SM data testing on various anti-influencer techniques like influencer injection.

F. Attempt *forecasting epidemic outbreaks* and evaluate post outbreak data.

G. E*valuate* the outcome of the intervention and the forecast.

Regarding our next step, it is evident from the studies that at least for prediction on virus outbursts, data alone cannot be enough. We are working on improving a forecasting algorithm in WEKA by incorporating other factors, such as real-life situations, that could trigger those outbursts.

References

[1] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.

[2] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

[3] Yu, S., & Kak, S. (2012). A survey of prediction using social media. arXiv preprint arXiv:1203.1647.

[4] Yang, Y., Gu, J., & Zhou, Z. (2016). Credit risk evaluation based on social media. Environmental research, 148, 582-585.

[5] Karppi, T., & Crawford, K. (2016). Social media, financial algorithms and the hack crash. Theory, Culture & Society, 33(1), 73-92.

[6] Dong, W., Liao, S., Xu, Y., & Feng, X. (2016). Leading Effect of Social Media for Financial Fraud Disclosure: A Text Mining Based Analytics.

[7] Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. International Review of Financial Analysis, 48, 272-281.

[8] Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down Wall Street with a tablet: A survey of stock market predictions using the web. Journal of Economic Surveys, 30(2), 356-369.

[9] Abbasi, O. R., Alesheikh, A. A., & Sharif, M. (2017). Ranking the City: The Role of Location-Based Social Media Check-Ins in Collective Human Mobility Prediction. ISPRS International Journal of Geo-Information, 6(5), 136.

[10] Zamani, M., & Schwartz, H. A. (2017). Using Twitter Language to Predict the Real Estate Market. EACL 2017, 28.

[11] Guidi, M. (2017). Extracting Information from Social Media to Track Financial Markets (Bachelor's thesis, Università Ca'Foscari Venezia).

[12] Barakos, M. (2015). Social Media and Forecasting: What is the Potential of Social Media as a Forecasting Tool? (Bachelor's thesis, University of Twente). Retrieved from on September 17th, 2017: http://essay.utwente.nl/67324/1/Barakos_BA_MB.pdf.

[13] Chen, J., Chen, H., Wu, Z., Hu, D., & Pan, J. Z. (2017). Forecasting smog-related health hazard based on social media and physical sensor. Information Systems, 64, 281-291.

[14] Doub, A. E., Small, M., & Birch, L. L. (2016). A call for research exploring social media influences on mothers' child feeding practices and childhood obesity risk. Appetite, 99, 298-305.

[15] Finch, K. C., Snook, K. R., Duke, C. H., Fu, K. W., Tse, Z. T. H., Adhikari, A., & Fung, I. C. H. (2016). Public health implications of social media use during natural disasters, environmental disasters, and other environmental concerns. Natural Hazards, 83(1), 729-760.

[16] Pekar, V., & Binner, J. (2017). Forecasting Consumer Spending from Purchase Intentions Expressed on Social Media. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 92-101).

[17] Jin, F., Wang, W., Chakraborty, P., Self, N., Chen, F., & Ramakrishnan, N. (2017). Tracking Multiple Social Media for Stock Market Event Prediction. In *Industrial Conference on Data Mining* (pp. 16-30). Springer, Cham.

[18] Chaffey, D. (2016). Global social media research summary 2016. Smart Insights: Social Media Marketing.

[19] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques* (p.32). Elsevier.

[20] Kim, D., Kim, J. H., & Nam, Y. (2014). How does industry use social networking sites? An analysis of corporate dialogic uses of Facebook, Twitter, YouTube, and LinkedIn by industry type. Quality & Quantity, 48(5), 2605-2614.

[21] Felix, R., Rauschnabel, P. A., & Hinsch, C. (2017). Elements of strategic social media marketing: A holistic framework. Journal of Business Research, 70, 118-126.

[22] Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *Signal Processing, Communication, Power and Embedded System (SCOPES), 2016 International Conference on* (pp. 1345-1350). IEEE.

[23] Karami, A., Bennett, L. S., & He, X. (2018). Mining Public Opinion About Economic Issues: Twitter and the US Presidential Election. *arXiv preprint arXiv:1802.01786*.

[24] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press (p.7).

[25] Candan, K. S., & Sapino, M. L. (2010). *Data management for multimedia retrieval*. Cambridge University Press (p.14).

[26] Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, 198-211.

[27] Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *3*(2), 1-159.

[28] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 492-499). IEEE Computer Society.

[29] Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*, *23*(5), 528-543.

[30] Sunstein, C. R. (2018). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

[31] Cioffi-Revilla, C. (2017). Computation and social science. In *Introduction to computational social science* (pp. 35-102). Springer, Cham.

[32] Gil de Zuniga, H., & Diehl, T. (2017). Citizenship, social media, and big data: Current and future research in the social sciences. *Social Science Computer Review*, *35*(1), 3-9.

[33] De Choudhury, M., Sharma, S., & Kiciman, E. (2016). Characterizing dietary choices, nutrition, and language in food deserts via social media. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (pp. 1157-1170). ACM.

[34] Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city.

[35] Tomeny, T. S., Vargo, C. J., & El-Toukhy, S. (2017). Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009-15. Social Science & Medicine.

[36] Aquino, F., Donzelli, G., De Franco, E., Privitera, G., Lopalco, P. L., & Carducci, A. (2017). The web and public confidence in MMR vaccination in Italy. Vaccine, 35(35), 4494-4498.

[37] Dredze, M., Broniatowski, D. A., Smith, M., & Hilyard, K. M. (2016). Understanding vaccine refusal: why we need social media now. American journal of preventive medicine, 50(4), 550.

[38] Radzikowski, J., Stefanidis, A., Jacobsen, K. H., Croitoru, A., Crooks, A., & Delamater, P. L. (2016). The measles vaccination narrative in Twitter: a quantitative analysis. JMIR public health and surveillance, 2(1).

[39] Ruizendaal, R. (2016). The Predictive Power of Social Media: Using Twitter to predict Spotify streams for newly released music albums (Master's thesis, University of Twente).

[40] Schade, L. (2015). Social Media and Forecasting: What is the Predictive Power of Social Media (Bachelor's thesis, University of Twente).

[41] Lassen, N. B., la Cour, L., & Vatrapu, R. (2017). Predictive Analytics with Social Media Data. The SAGE Handbook of Social Media Research Methods, 328.

[42] Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. Science advances, 2(3), e1500779.

[43] McDonald, R., & Mao, X. (2015). Forecasting the 2015 General Election with Internet Big Data: An Application of the TRUST Framework (No. 2016_03).

[44] Vijayaraghavan, R., Garikipati, R., Chang, A., & Kannan, P. V. (2016). *U.S. Patent No. 9,519,936*. Washington, DC: U.S. Patent and Trademark Office.

[45] Kim, J., Cha, M., & Lee, J. G. (2017). Nowcasting commodity prices using social media. PeerJ Computer Science, 3, e126.

[46] Kalmer, N.P. (2015) the predictive power of Social Media Analytics: To what extent can SM Analytics techniques be classified as reliable and valid predictive tools? Retrieved from on September 17th, 2017: http://essay.utwente.nl/68516/1/Kalmer_BA_MB.pdf.

[47] McClellan, C., Ali, M. M., Mutter, R., Kroutil, L., & Landwehr, J. (2017). Using social media to monitor mental health discussions− evidence from Twitter. Journal of the American Medical Informatics Association, 24(3), 496-502.

[48] Kim, W. G., Li, J. J., & Brymer, R. A. (2016). The impact of social media reviews on restaurant performance: The moderating role of excellence certificate. International Journal of Hospitality Management, 55, 41-51.

[49] Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory? Using Twitter to predict the UK 2015 General Election. Electoral Studies, 41, 230-233.

[50] Isotalo, V., Saari, P., Paasivaara, M., Steineker, A., & Gloor, P. A. (2016). Predicting 2016 US Presidential Election Polls with Online and Media Variables. In Designing Networks for Innovation and Improvisation (pp. 45-53). Springer International Publishing.

[51] Cameron, M. P., Barrett, P., & Stewardson, B. (2016). Can social media predict election results? Evidence from New Zealand. Journal of Political Marketing, 15(4), 416-432.

[52] Ni, M., He, Q., & Gao, J. (2017). Forecasting the subway passenger flow under event occurrences with social media. IEEE Transactions on Intelligent Transportation Systems, 18(6), 1623-1632.

[53] Sadilek, A., Kautz, H. A., DiPrete, L., Labus, B., Portman, E., Teitel, J., & Silenzio, V. (2016). Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. In AAAI (pp. 3982-3990).

[54] Phillips, L., Dowling, C., Shaffer, K., Hodas, N., & Volkova, S. (2017). Using Social Media to Predict the Future: A Systematic Literature Review. arXiv preprint arXiv:1706.06134.

[55] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th int'l conf. on World wide web* (pp. 851-860). ACM.

[56] Lin, Z., Jin, H., Robinson, B., & Lin, X. (2016). Towards an accurate social media disaster event detection system based on deep learning and semantic representation.

[57] Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2017). Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, 0165551517698298.

[58] Hudson, S., Huang, L., Roth, M. S., & Madden, T. J. (2016). The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors. *International Journal of Research in Marketing*, *33*(1), 27-41.

[59] Oh, O., Kwon, K. H., & Rao, H. R. (2010, August). An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010. In *ICIS* (Vol. 231).

[60] Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. *SBP*, *12*, 231-238.

[61] Pritam Gundecha and Huan. Liu, Mining social media: a brief introduction, Tutorials in Operations Research 1 (2012).

[62] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, *61*(12), 2544-2558.

[63] Pew Research Center September (2012, September 9), 'Section 1: The Obama-Romney Matchup'. Retrieved March 18, 2018, from http://www.people-press.org/2012/09/19/section-1-the-obama-romneymatchup/

[64] Gayo-Avello, D. (2012). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441*.

[65] Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A. and Chakraborty, B., 2016. A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal*.

[66] Nguyen, T.H., Shirai, K. and Velcin, J., 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), pp.9603-9611.

[67] Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O. and Brownstein, J.S., 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, *11*(10), p. e1004513.

[68] Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A. and Chakraborty, B., 2018. A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal*, *9*(3), pp.365-378.

[69] Subramani, S., Michalska, S., Wang, H., Whittaker, F., & Heyward, B. (2018, October). Text Mining and Real-Time Analytics of Twitter Data: A Case Study of Australian Hay Fever Prediction. In *International Conference on Health Information Science* (pp. 134-145). Springer, Cham.

[70] Gruner, R. L., Vomberg, A., Homburg, C., & Lukas, B. A. (2018). Supporting new product launches with social media communication and online advertising: sales volume and profit implications. *Journal of Product Innovation Management*.

[71] ElTayeby, O., Eaglin, T., Abdullah, M., Burlinson, D., Dou, W., & Yao, L. (2018). A feasibility study on identifying drinking-related contents in Facebook through mining heterogeneous data. *Health informatics journal*, 1460458218798084.

[72] Gaikar, D. D., Marakarkandy, B., & Dasgupta, C. (2015). Using Twitter data to predict the performance of Bollywood movies. *Industrial Management & Data Systems*, *115*(9), 1604-1621.

[73] He, W., Shen, J., Tian, X., Li, Y., Akula, V., Yan, G., & Tao, R. (2015). Gaining competitive intelligence from social media data: Evidence from two largest retail chains in the world. *Industrial management & data systems*, *115*(9), 1622-1636.

[74] Ong, Y. X., & Ito, N. (2019). "I Want to Go There Too!" Evaluating Social Media Influencer Marketing Effectiveness: A Case Study of Hokkaido's DMO. In *Information and Communication Technologies in Tourism 2019* (pp. 132-144). Springer, Cham.

[75] Zhan, Y., Etter, J. F., Leischow, S., & Zeng, D. (2018). Electronic cigarette usage patterns: a case study combining survey and social media data. *Journal of the American Medical Informatics Association*, *26*(1), 9-18.

[76] Su, Y., Li, X., Tang, W., Xiang, J., & He, Y. (2018, June). Next check-in location prediction via footprints and friendship on location-based social networks. In *2018 19th IEEE Int'l Conf. on Mobile Data Management (MDM)* (pp. 251-256). IEEE.

[77] Gao, Y., Zhong, Y., Preotiuc-Pietro, D., & Li, J. J. (2019). Predicting and Analyzing Language Specificity in Social Media Posts.

[78] Ricard, B. J., Marsch, L. A., Crosier, B., & Hassanpour, S. (2018). Exploring the Utility of Community-Generated Social Media Content for Detecting Depression: An Analytical Study on Instagram. *Journal of medical Internet research*, *20*(12).

[79] Baghel, N., Kumar, Y., Nanda, P., Shah, R. R., Mahata, D., & Zimmermann, R. (2018). Kiki Kills: Identifying Dangerous Challenge Videos from Social Media. *arXiv preprint arXiv:1812.00399*.

[80] Steinert, L., & Herff, C. (2018). Predicting altcoin returns using social media. *PloS one*, *13*(12), e0208119.

[81] Matta, M., Lunesu, I., & Marchesi, M. (2015, June). Bitcoin Spread Prediction Using Social and Web Search Media. In *UMAP Workshops*.

[82] Rossi, C., Acerbo, F. S., Ylinen, K., Juga, I., Nurmi, P., Bosca, A., ... & Alikadic, A. (2018). Early detection and information extraction for weather-induced floods using social media streams. *Int'l Journal of Disaster Risk Reduction*.

[83] L. Oikonomou and C. Tjortjis, 'A Method for Predicting the Winner of the USA Presidential Elections using Data Extracted from Twitter', 3rd IEEE SEEDA-CECNSM18), 2018.

[84] Hristidis, V., Chen, S. C., Li, T., Luis, S., & Deng, Y. (2010). Survey of data management and analysis in disaster situations. *Journal of Systems and Software*, *83*(10), 1701-1714.

[85] Borges, P.V.K., and Izquierdo E., "A probabilistic approach for vision-based fire detection in videos." *IEEE transactions on circuits and systems for video technology* 20.5 (2010): 721-731.

[86] Chen, J., Ying, P. & Zou, M., Improving music recommendation by incorporating social influence. Multimed Tools Appl (2019) 78: 2667.

[87] Chuman, T., Iida, K., & Kiya, H. (2017, December). Image manipulation on social media for encryption-then-compression systems. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 858-863). IEEE.

[88] Ning, J., Singh, I., Madhyastha, H. V., Krishnamurthy, S. V., Cao, G., & Mohapatra, P. (2014, October). Secret message sharing using online social media. In *2014 IEEE Conference on Communications and Network Security* (pp. 319-327). IEEE.

[89] Muhammad, K., Ahmad, J., Rho, S., & Baik, S. W. (2017). Image steganography for authenticity of visual contents in social networks. *Multimedia Tools and Applications*, *76*(18), 18985-19004.

[90] Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, *6*(2), 9.

[91] Huh, J. H., & Seo, K. (2014). Development of competency-oriented social multimedia computer network curriculum. *Journal of Multimedia Information System*, *1*(2), 133-142.

[92] Huh, J. H. (2018). Big data analysis for personalized health activities: machine learning processing for automatic keyword extraction approach. *Symmetry*, *10*(4), 93.

[93] Vanden Abeele, V. A., & Van Rompaey, V. (2006, April). Introducing human-centered research to game design: designing game concepts for and with senior citizens. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 1469-1474). ACM.

[94] McGee, J., Caverlee, J., & Cheng, Z. (2013, October). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 459-468). ACM.

[95] Tang, J., Chang, S., Aggarwal, C., & Liu, H. (2015, February). Negative link prediction in social media. In *Proceedings of the eighth ACM international conference on web search and data mining* (pp. 87-96). ACM.

[96] Liao, X., Yu, Y., Li, B., Li, Z., & Qin, Z. (2019). A new payload partition strategy in color image steganography. *IEEE Transactions on Circuits and Systems for Video Technology*.

[97] Liao, X., Li, K., & Yin, J. (2017). Separable data hiding in encrypted image based on compressive sensing and discrete fourier transform. *Multimedia Tools and Applications*, *76*(20), 20739-20753.

[98] Liao, X., Qin, Z., & Ding, L. (2017). Data embedding in digital images using critical functions. *Signal Processing: Image Communication*, *58*, 146-156.

[99] Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2013, July). Opinion mining on the web 2.0–characteristics of user generated content and their impacts. In *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 35-46). Springer, Berlin, Heidelberg.

[100] Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S. M., Schaller, S., & Holzinger, A. (2012, December). On text preprocessing for opinion mining outside of laboratory environments. In *International Conference on Active Media Technology* (pp. 618-629). Springer, Berlin, Heidelberg.

[101] Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2015). Computational approaches for mining user's opinions on the Web 2.0. *Information Processing & Management*, *51*(4), 510-519.

[102] Koukaras, P., Tjortjis, C., & Rousidis, D. (2019). Social Media Types: introducing a data driven taxonomy. *Computing*, 1-46.

[103] Siddiqui, S., & Singh, T. (2016). Social media its impact with positive and negative aspects. *International Journal of Computer Applications Technology and Research*, *5*(2), 71-75.

[104] Abdulahi, A., Samadi, B., & Gharleghi, B. (2014). A study on the negative effects of social networking sites such as facebook among asia pacific university scholars in Malaysia. *International Journal of Business and Social Science*, *5*(10).

[105] Seabrook, E. M., Kern, M. L., & Rickard, N. S. (2016). Social networking sites, depression, and anxiety: a systematic review. *JMIR mental health*, *3*(4), e50.

[106] Tsihrintzis, G. A., Virvou, M., Sakkopoulos, E., & Jain, L. C. (2019). Applications of Learning and Analytics in Intelligent Systems. In *Machine Learning Paradigms* (pp. 1-6). Springer, Cham.

[107] Guedes, E., Nardi, A. E., Guimarães, F. M. C. L., Machado, S., & King, A. L. S. (2016). Social networking, a new online addiction: a review of Facebook and other addiction disorders. *MedicalExpress*, *3*(1).