

Sentiment analysis in SemEval: a review of sentiment identification approaches

Bousselham El Haddaoui, Raddouane Chiheb, Rdouan Faizi, Abdellatif El Afia

National Higher School for Computer Science and Systems Analysis, Mohammed V University in Rabat, Rabat, Morocco

Article Info

Article history:

Received Jun 8, 2022

Revised Jul 16, 2022

Accepted Aug 18, 2022

Keywords:

Deep learning

Machine learning

Sentiment analysis

Social media

Transformers

ABSTRACT

Social media platforms are becoming the foundations of social interactions including messaging and opinion expression. In this regard, sentiment analysis techniques focus on providing solutions to ensure the retrieval and analysis of generated data including sentiments, emotions, and discussed topics. International competitions such as the International Workshop on Semantic Evaluation (SemEval) have attracted many researchers and practitioners with a special research interest in building sentiment analysis systems. In our work, we study top-ranking systems for each SemEval edition during the 2013-2021 period, a total of 658 teams participated in these editions with increasing interest over years. We analyze the proposed systems marking the evolution of research trends with a focus on the main components of sentiment analysis systems including data acquisition, preprocessing, and classification. Our study shows an active use of preprocessing techniques, an evolution of features engineering and word representation from lexicon-based approaches to word embeddings, and the dominance of neural networks and transformers over the classification phase fostering the use of ready-to-use models. Moreover, we provide researchers with insights based on experimented systems which will allow rapid prototyping of new systems and help practitioners build for future SemEval editions.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Bousselham El Haddaoui

National Higher School for Computer Science and Systems Analysis, Mohammed V University

Rabat, Morocco

Email: bousselham.haddaoui@um5s.net.ma

1. INTRODUCTION

Technological growth has shaped, by different means, human interactions. Social media platforms, forums, and news websites are providing digital services allowing internet users to perform daily activities such as information sharing, messaging, and public discussions. With a 4.65B active social media users representing 58.7% of the world population estimated at 7.89B [1], industry stakeholders (i.e., e-commerce, politics, and healthcare) show interest in the study of the large generated data from user interactions to extract insights and business value for decision-making, urging the need for the design of suitable techniques to process the available data with reasonable time and cost investment considerations. In this regard, sentiment analysis, which is an interdisciplinary field that focuses on the extraction and analysis of information from diverse data sources (i.e., short and large text, images, and videos), along with natural language processing (NLP), computational linguistics, and text mining techniques have provided the foundations to design automated and industry-grade systems to address these challenges [2]. Sentiment analysis focuses on various tasks including sentiment classification, aspect extraction, and topic modeling. It has evolved from a simple classification task to a problem

of automatic opinion detection and opinion-related attributes identification [3] profiting from advancements in neural networks and other emerging research trends such as pre-trained language models (PLM) [4] and neuro-symbolic artificial intelligence (AI) [5]. Currently, sentiment analysis systems are widely used in the social media context and have many applications in other domains such as healthcare, social sciences, market research, and political science [6].

Research studies, in this field, usually cover one or two aspects (i.e., word representations, features extraction, and classification) which result in knowledge sparsity. The need for automated systems motivated a research interest shifting toward the sentiment analysis pipeline's design and implementation, thus providing state-of-the-art baselines. In this respect, international competitions such as the International Workshop on Semantic Evaluation (SemEval) are committed to raising awareness about sentiment analysis in the industry and academic communities. They have focused on the study of various social phenomena through their editions. SemEval organizes several sub-tasks covering trending sentiment analysis topics such as sentiment and emotion detection and classification, social phenomena such as hate speech, sarcasm, and offense, named entity recognition, and ensures the evaluation of the submitted systems. Furthermore, the submitted systems are ready-to-use technical implementations of sentiment analysis systems that are based on recent research trends, thus providing a variety of system choices and fostering the improvement of previously submitted systems.

In our study, we aim to provide an in-depth analysis of submitted systems through the SemEval editions marking the evolution of used techniques on various sentiment analysis aspects, highlight the opportunities and limitations of proposed approaches, and present future research trends of sentiment analysis systems implementations. In this respect, we review the five top-performing systems for each edition from 2013 to 2021 to provide insights regarding the evolution of used datasets, their preparation, and the annotation process. Moreover, we present the main used techniques and research innovations concerning various sentiment analysis systems aspects including data acquisition, data preparation, preprocessing techniques, features engineering, and classification approaches.

The rest of the paper is structured section 2 presents the history of the SemEval competition and a description of the challenges and research questions for each edition. Section 3 provides a review and analysis of the studied systems focusing on defined aspects such as datasets, preprocessing techniques, word representation and features engineering, and classification models. A timeline tracking the evolution of used techniques for each aspect was also provided highlighting key contributions to the sentiment analysis field. In section 4, key findings are presented and discussed. Conclusions are presented in section 5.

2. HISTORY OF SEMEVAL TASKS

The SemEval competitions focus on various NLP research topics including text similarity and question answering, time and space, word sense disambiguation and induction, learning semantic relations, and sentiment analysis. In our study, we considered the sentiment analysis special track and narrowed our coverage to the sentence classification task. The timeline of the covered SemEval editions, the subject of our study, is presented below.

2.1. SemEval-2013 task 2

In [7], the first edition of the competition, the subtasks focused on the contextual polarity disambiguation and the message polarity classification. The key challenges for these tasks included the lack of suitable datasets for training and submissions assessment, the informal nature and the creative content retrieved from Twitter, and the overall sentiment conveyed in messages containing opposite sentiments. A total of 44 teams participated in the task.

2.2. SemEval-2014 task 9

In the second edition [8], the organizers deepened the research questions for the previous subtasks by introducing two additional datasets covering formal content and sarcasm social content. The main challenges for these tasks included new content patterns such as the use of creative spelling and punctuation, slang, new words, and abbreviations. Sarcasm handling was the key addressed topic along with defining sentiment strength in tweets conveying opposite sentiments. A total of 46 teams participated in the task.

2.3. SemEval-2015 task 10

The 3rd edition of the competition [9] introduced new challenges related to sentiment analysis. In addition to the previous two subtasks, message sentiment and overall trend toward a topic were investigated along with the strength of association of specific words and positive sentiment. Challenges for these tasks included new patterns in social content such as emoticons, acronyms, and poor grammatical structure. A total of 40 teams participated in the task.

2.4. SemEval-2016 task 4

In [10], the 4th edition of the competition, the sub-tasks introduced the ordinal classification instead of the usual binary classification for the previous message polarity and sentiment toward topic tasks, and the quantification task which implies the study of the distribution of classes in unlabeled datasets. The key challenges for these tasks are the lack of training datasets and class imbalance, the difficulty of handling multi-class ordinal classifications, and the evaluation metrics for the quantification task. A total of 43 teams participated in the task.

2.5. SemEval-2017 task 4

The 5th edition [11] was a rerun of the previous competition, the organizers initiated the cross-language sentiment analysis with the introduction of the Arabic language for all the subtasks. In addition, the datasets provided user-related demographic information such as age, gender, and location. for usage as extra features. The main challenges for these tasks involved the lack of training data for the Arabic language, the high featured level of the language, and the abundant use of dialect forms and spelling variants. Besides, the cross-language nature of the ordinal classification and topic-related sentiment extraction. A total of 48 teams participated in the task.

2.6. SemEval-2018 task 1

In [12], the 6th edition of the competition, the organizers focused on inferring the mental state of the user. In this respect, the subtasks covered the emotional intensity and the valence ordinal classification and regression, and emotion classification in a cross-language context. The limited training datasets and the cross-language were the main challenges in the preparation for this task. A total of 75 teams participated in the task marking an increased interest in the competition.

2.7. SemEval-2019 task 9

In the 7th edition [13], the competition focused on offensive language identification and categorization. The proposed tasks included offensive language detection, categorization, and offense target identification. Challenges for these tasks included the process of building the evaluation dataset which should overcome the annotator's bias and an extensive understanding of offensive language categorization. A total of 155 teams participated in the task showing a growing interest in the task.

2.8. SemEval-2020 task 12

In [14], the 8th edition of the competition, the organizers introduced a multilingual aspect for offense identification and categorization. The provided dataset followed the best practices in abusive language collection [15] and covered five languages which are Arabic, Danish, English, Greek, and Turkish. Besides the challenges from the previous edition, the targeted offense presented many challenges including implicit and explicit offenses for individual and group targets that could be based on gender and religious beliefs or ethnicity. A total of 145 teams participated in the task.

2.9. SemEval-2021 task 7

In the 9th edition [16], the scope of the competition investigated both humor and offense through 4 subtasks which are humor detection, humor rating and controversy prediction, and offense detection. The main challenges in this competition included the ability to differentiate between humor and offense in tweets, and the perception of humor that can vary depending on age, gender, and personality [17]. Moreover, the levels of controversy in judgments between interceptors were also a challenge to address in these subtasks. A total of 62 teams participated in the task, the competition organized 11 different tasks.

SemEval competitions addressed many active sentiment analysis topics through their tasks, starting from the basic sentence and message polarity, message polarity toward a topic, to more advanced topics such

as the study of figures of speech (i.e. sarcasm, offense, and humor), cross-language and ordinal classification. Moreover, notable limitations were encountered including the difficulty of preparing suitable datasets for training and evaluation purposes and the specificity of social media language. The proposed systems, in the previous competitions, provided ready-to-use solutions to answer the research questions marking a timeline of the evolution of tools and methodologies used in the sentiment analysis research area.

3. PROPOSED SYSTEMS

3.1. Datasets

Since 2013, the competition has focused mainly on datasets collected from Twitter [18]. Alternative sources were considered including Reddit, News, and Kaggle to enrich the main dataset depending on the chosen topic. Data collection and annotation processes are required to prepare the competition datasets for training and evaluation purposes. The data collection phase is ensured by endpoints that are provided, by social media platforms, for content retrieval using open source software, otherwise scraping techniques or ready-to-use datasets are considered. Following this, organizers pull out content with no sentiment-bearing words to reduce class imbalance by keeping content with a score superior to 0.3 using SentiWordNet [19] in the 2013-2018 editions and remove duplicates relying on bag of words (BOW) cosine similarity that exceeds 0.6 in 2018 edition.

The data annotation process is conducted after collection and cleaning depending on the task nature, various methodologies were considered through the editions including crowdsourcing platforms such as Amazon's Mechanical Turk [20], Crowdfunder (rebranded to Appen [21]), and Prolific [22]. Furthermore, majority voting systems were used based on expert manual annotations in [13]. An initiative to use semi-supervised annotation systems based on ensemble techniques was introduced in [14], outputs from selected voters including pointwise mutual information (PMI) [23], fast text [24], long short-term memory (LSTM) networks [25], and bidirectional encoder representations from transformers (BERT) [26], are combined to assess the agreement score between voting models using Fleiss' K inter-annotator agreement (IAA) [27].

The manual aspect of the preparation of the datasets makes it extremely slow and very expensive, thus the output of this process results usually in small to medium size datasets. The lack of suitable, class imbalance, and insufficient training datasets was the main restriction for the early editions considering the dependence of proposed systems on linguistic resources [28] and the limited datasets improvement which does not impact the overall score during the evaluation process [8]. The prepared datasets size depends on the nature of the competition topics, and available datasets are reused and enriched by organizers for the reconducted or similar tasks as detailed in Table 1. Moreover, new datasets are prepared for specific tasks such as humor, offense, racism, and for which the sentiment distribution is not described since organizers used different classes (i.e., offensive/not offensive, humor/not humor, emotions ordinal classification). As for the datasets evolution, the organizers collected 17,401 tweets in 2013 to reach 63,677 tweets in 2017 as presented in Figure 1, while new datasets were manually prepared for specific tasks from 2018 to 2021. The SOLID dataset (9,093,037 tweets) used in the 2020 edition [14] which was prepared using a semi-supervised technique [29] was not considered in Figure 1 to appreciate the evolution of original datasets, manually collected, and annotated by experts. To overcome the limitations of size and quality in the competitions datasets, participants further enriched the training data by introducing public datasets from known repositories such as Kaggle [30], UCI [31], GitHub [32], and collected social media content from social platforms using topic related word seeds [16] or emoticons [33] as keywords for queries.

3.2. Data preprocessing

Social media content, the primary focus of SemEval competitions, gained big data attributes such as volume, variety, and velocity over time. Furthermore, the rich nature and shape of social messages which hold information about individuals and their interactions [7], and convey user sentiment [34] can be altered by different means considering the creative and informal social content. Misspellings, poor grammatical structure, hashtags, punctuation, new words, emoticons, acronyms, and slang [9] are language phenomena qualified as noise that causes extreme lexical sparsity [35]. To handle these language issues, preprocessing tasks had become an essential component of every sentiment analysis system. In addition, other complex methods such as word context disambiguation, and out-of-vocabulary (OOV) handling can be applied at this level to improve content quality and system performance [36].

The need for automation for sentiment analysis systems [28] gives grounds for advanced use of preprocessing techniques and raises research interest in their usability and effectiveness in the context of social media. The need to clean noisy data [7] without affecting text meaning is the main concern of preprocessing, especially for social media content whose features are linguistically driven and require specific text processing [37]. Moreover, text normalization techniques including hashtags segmentation, emojis conversion, and spell correction [14] have shown their effectiveness in dealing with large vocabularies [38], negation handling [39], and OOVs [40]. The effectiveness of some preprocessing techniques such as punctuation removal is still controversial since punctuation may not affect the general meaning, while in other cases (i.e., multiple exclamations), might be useful as additional features [41].

Table 1. SemEval 2013-2021 datasets statistics

Edition	Dataset size	Data source	Sentiment distribution		
			Positive	Negative	Neutral
2021	8,000	Twitter			
	2,000	Kaggle			
2020	9,107,137	Twitter			
2019	14,100	Twitter			
2018	26,184	Twitter			
2017	62,617	Twitter	22,277	11,812	28,528
	50,158	Twitter	19,855	9,004	21,299
2016	2,093	SMS	492	394	1,207
	1,142	Live journal	427	304	411
2015	19,526	Twitter	7,864	3,014	8,648
	2,093	SMS	492	394	1,207
2014	1,142	Live journal	427	304	411
	17,134	Twitter	6,824	2,649	7,661
2013	2,093	SMS	492	394	1,207
	1,142	Live journal	427	304	411
2013	15,196	Twitter	5,810	2,407	6,979
	2,094	SMS	492	394	1,208

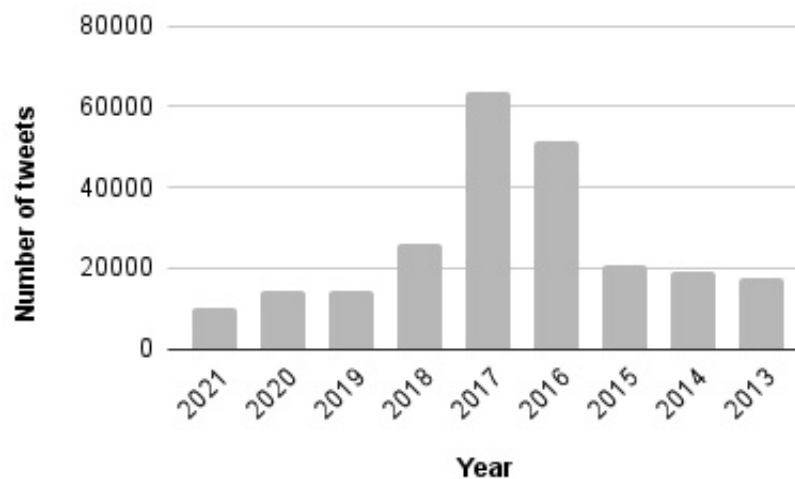


Figure 1. The evolution of datasets size during the SemEval competition

In our study scope, top-performing participants combined at least one to four preprocessing techniques as shown in Figure 2, the appeal to preprocessing techniques is usually motivated by a will to improve state-of-the-art models. Whereas a decline in the use of preprocessing refers to the introduction of new models or methodology. There was a similarity in used processing techniques between the 2013 and 2014 editions of the competition [8], while no new trends were observed in the 2016 edition [10]. A particular use of ready-to-use text preprocessors such as the Ark Tokenizer [42] and the NLTK TweetTokenizer [43], Stanford Core NLP [44],

and Keras [45] by participants was noted in the 2019 edition with a focus on the combination of preprocessing techniques in the 2021 edition. The review of used techniques through the considered systems allowed us to perform a summarization by usage frequency, below is a list of used preprocessing techniques:

- High usage frequency: tokenization, lowercasing, uniform resource locator (URL), user mentions, and special characters (Unicode and XML) removal.
- Average usage frequency: parts of speech (POS), convert emojis to words, duplicates, punctuation, and hashtag removal.
- Low usage frequency: lemmatization, negation handling, spelling correction, stemming, truncate tweet, remove stop words, and extra white space.

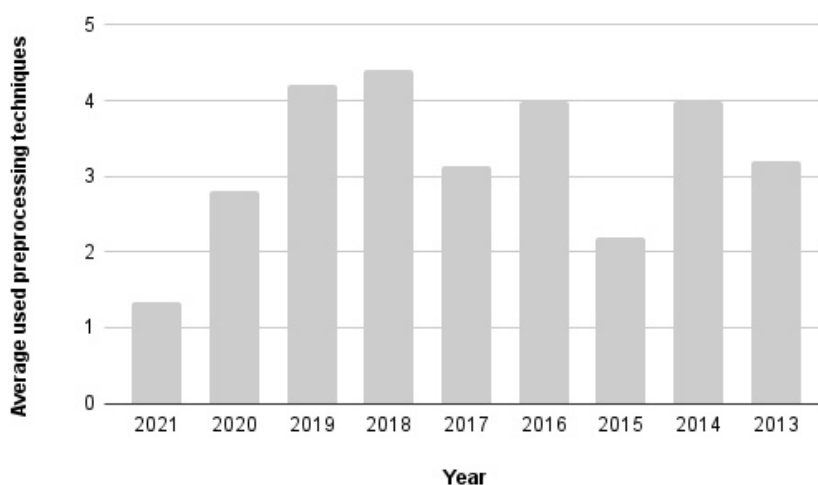


Figure 2. The Average used preprocessing techniques by edition

The usage of preprocessing techniques depends on the analysis approach, some techniques remove useless content while others improve classifier performance [46]. The studied systems used various approaches such as machine learning classifiers, neural networks, and transformers. Figure 3 shows usage ratios of preprocessing techniques by classification approach, transformer-based approaches rely less on preprocessing techniques compared to other approaches which consider it a key component of every sentiment analysis pipeline.

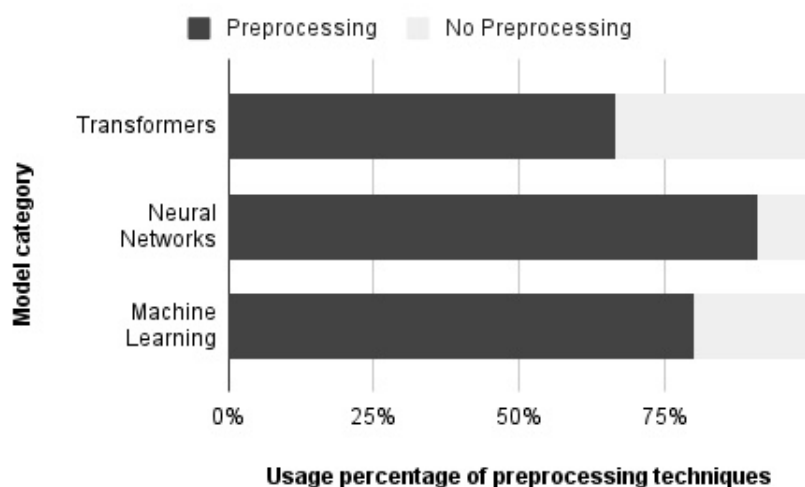


Figure 3. Usage of preprocessing techniques per sentiment analysis model category

3.3. Data representation and features engineering

Sentiment can be expressed in different forms on social media, and in most cases in a discrete [47] and an implicit way [48] which makes the identification and extraction tasks difficult to perform. The complexity of human sentiment can vary from basic emotions, described in the basic emotion model, such as joy, sadness, and fear that can be physiologically and cognitively expressed [49], [50], to more complex ones studied in the valence-arousal-dominance (VAD) model [51] which categorize emotions into a dimensional grouping. Valence (positiveness-negativeness), arousal (active-passive), and dominance (dominant, submissive) are dimensions of the emotion space where sentiments can be represented. Furthermore, a sentiment representation was proposed in [48] into a machine-understandable form consisting of an entity $E(j)$, the aspect of the entity $A(j, k)$, the sentiment $S(j, k, i, l)$, the sentiment holder $H(i)$, and the time $T(l)$. In addition, the geographical location was incorporated to enrich the representation in [52].

State-of-the-art systems depend heavily on linguistic resources, extensive features engineering, and tuning [28]. Moreover, sentiment categorization relies on feature choice which remains a primary challenge for the sentiment analysis systems [37], and feature selection strategies (i.e., linguistic, lexical, or mixed strategy). In [7], most systems proposed a lexicon-based strategy while pointing to a lack of sentiment lexicons [53]. Available resources, as presented in [40], [53], [34], [8], [54], [55], [12], included formal lexicons such as General Inquirer [56], MPQA Subjectivity Lexicon [57], SentiWordNet [58], and informal ones such as AFINN-111 [59], Bing liu's opinion lexicon [60], NRC hashtag sentiment lexicon [61], Sentiment140 Lexicon [62]. Other lexicon datasets were used by participants including DeRose [63], Urban Dictionary slang dictionary, and large-vocabulary distributional semantic models (DSM) constructed from Wacky web-crawled large corpora [64] and the Google Web 1T 5-Grams database [65]. In addition, research efforts focused on improving proposed systems with manually created domain-dependent and independent taxonomies from open data sources such as forums, news, etc. which require in some cases domain knowledge [66] or based on sentiment-bearing word seeds including hashtags, emoticons, keywords, and from social media platforms [53].

Although lexicon-based approaches contribute effectively to determining the overall sentiment, additional feature engineering may be required to further improve model performance. In this respect, specific features of social context are considered including Twitter profile's demographic information such as user age, gender, location, and followers count) [11], other approaches are considered such as word sense disambiguation [40], negation handling [39], and emojis conversion [67]. The focus on features engineering is motivated by the contribution of rich features set on the classification models [34], and feature weighting may reflect the influence of each feature on the overall sentiment [35].

In [12], three main feature engineering techniques were highlighted including lexicon features, word n-grams, and word embeddings. Although important features are extracted from lexicons [9], lexicon-based approaches present many limitations resulting essentially from manual annotation effort, domain-dependent features crafted from sentence words, emotions, slang, and hashtags, [41] which motivated word-based features engineering. In word-based approaches, a variety of techniques were used by participants in [7] such as word-based (i.e., stems, words, clustering, n-grams), word-shape (i.e., capitalization, punctuation), syntactic (i.e., dependency relations, part of speech tags (POS)), and Twitter-specific features (i.e., emoticons, repeated letters, hashtags, URLs, abbreviations, and slang). Moreover, surface form features can be also considered including the number of elongated words, the number of hashtags [53], expanded words from a predefined word list [66], word shape, interjection (words that express a sentiment such as lol, hurrah) [34], the number of tokens in a sentence [41], negation and semantic features [54], all-caps and punctuation [68]. Another explored approach in this context is bag of words (BOW) [69], various sets of bags of words including unigrams, bigrams, and extended unigrams models were explored in [41]. The approach is one of the most important representation methods [70], and it becomes less effective in short texts and leads to increased data sparseness. In [38], additional techniques are provided to complement BOW with a denser representation including weighting relevant words in a BOW (BM25) [71], mapping words to clusters using brown clusters [72], or assigning weight to terms on each identified class using concise semantic analysis [73]. The performance yielded by lexicons and word-based approaches is coupled with a difficulty in relying on manual features, thus the need for a new approach [33].

Participants in recent SemEval editions focused on the latest research trends through the use of unsupervised learning of word embeddings [28]. Word2Vec [74] and global vectors (GloVe) [75] are two frequently used unsupervised word embedding techniques that provide general-purpose and multidimensional word embeddings and provide fine-tuning mechanisms for domain-dependent data representation [33]. Furthermore,

improvements to the previous techniques are used such as FastText [76] which enhances Word2Vec's ability to train on small datasets and generalize to unknown words, sentiment embeddings fusion, and sentiment-specific word embeddings that consider word sentiment in addition to syntactic and semantic contexts [77]. Moreover, character embeddings were explored to overcome the limitations of word embeddings in handling OOVs [36], while traditional approaches usually ignore OOVs (i.e., set to default or neutral words) ignoring the eventuality of being sentiment-bearing words. In the 2020 edition, systems relied more on contextualized word representations provided by transformers such as BERT [26], ELMo [78], RoBERTa [79], and multilingual BERT (mBERT), due to the significant improvements brought to various NLP tasks [80].

3.4. Classification models

SA systems give special consideration to the classification task as it represents the core component of the majority of proposed systems. Classification approaches fall, generally, under three categories including supervised, semi-supervised, and unsupervised techniques. The first editions of SemEval noted a trend focusing on supervised learning [9], models such as support vector machines (SVM), naive bayes (NB), maximum entropy, rule-based classifiers, and ensemble techniques [7] were the most used. Whereas, recently top-performing systems relied more on neural networks, language models, and features derived from existing emotion and sentiment lexicons [12]. Organizers paid special attention to constrained (relying only on provided datasets) and unconstrained systems (using additional resources) which allowed the use of transductive (the same task for different domains) and inductive (the same domain for different tasks) learning strategies. Four used model classes were identified during our study, categories include machine learning (ML), deep learning (DL), transformers, and ensemble models.

Initial approaches for classification were rule-based [7]. Rules matching may cover words and sequences, and are either handwritten, identified using statistical methods, or available software such as Synes-ketch [81]. Although they have high precision, rule-based systems require skilled linguists to define rules and lack generalization and scalability which are addressed and improved with machine learning algorithms [82]. Machine learning models provided state-of-the-art results during the 2013-2016 editions, SVM, maximum entropy, conditional random fields (CRFs), linear regression, and logistic regression were the most commonly used models [9]. Models were enhanced using cross-validation on training datasets [53], using loss functions such as hinge loss function to improve accuracy, and regularization such as L1, L2, or elastic net regularization (combination of L1 and L2) [37] to avoid overfitting. ML-based approaches have known limitations related to poor transfer learning and limited ability to learn complex patterns which motivated the use of DL approaches in a quest to further improve systems performance.

Since SemEval 2015, top-performing systems have used mainly DL models built using deep neural networks and word embeddings [9]. The limitation of transfer learning in ML models can be overcome by initializing neural networks with an embeddings layer pre-trained on available general-purpose or domain-specific corpus (i.e., GloVe 25d, 50d, 100d, and 300d, Google new corpus with 300d vectors) [83], networks can be also initialized and refined using distant learning as in [62] where a convolutional neural network (CNN) inspired from [84] was used for this purpose [85]. This usage trend was manifested in SemEval 2019 [13] with more than 70% of systems using various DL models including LSTM, bidirectional LSTM (BiLSTM), recurrent neural networks (RNN), CNN, and gated recurrent unit (GRU) neural networks. Although DL systems proved their high performance, representation learning techniques (based on DL neural networks) can benefit from manually engineered features [13]. Furthermore, additional optimization techniques were considered such as adding small perturbations on input samples using adversarial examples [86] which proved to improve models loss [87] along with advanced techniques that may be used including grid search, random search, and genetic algorithm to learn neural networks hyperparameters [88].

Driving a paradigm shift in the use of DL approaches, pre-trained language models (PLMs) or transformers are the new evolution of the neural network models. BERT [26] has shown significant performance and proved resistant to overfitting compared to other models [89], and it helps better understanding of sentence meaning and generates expressive word-level representations due to the inherent data noise in social media content [90]. Furthermore, PLM such as BERT and universal language model fine-tuning (ULMFiT) [91] provided better results in multilingual classification [16] compared to systems based on linear models and optimized using evolutionary algorithms. New PLMs were introduced that outperform existing models to optimize existing models in various directions such as providing lite versions that reduce parameters, increase models speed, and reduce memory consumption (i.e., ALBERT [92] for BERT). Other alternatives focused on training

strategies such as XLNet [93] which introduced an automatic regressive pre-training method [94]. This notable advancement in performance using transformers comes with many limitations including a negative impact of class imbalance [95], and the required intensive computation using large PLM models [96].

Ensemble learning was also considered the SemEval systems combining top-performing ML and DL models [97] provided a hybrid approach taking the benefits of each model category. The weighting scheme varies depending on the adopted strategy, soft voting [83] and hard majority voting [94] are known options that can be improved using algorithms (i.e., limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS)) to optimize weights attributed to each model.

During the 2013-2021 SemEval editions, the evolution of models used in sentiment analysis-related tasks was observed. Contributions evolved using ML and DL models to recent Transformers following research trends to improve the performance and accuracy of proposed systems. Figure 4 shows the evolution of usage per mode category backed by insights and taking into consideration the competition timeline and various constraints. The dominance of ML systems was challenged after the 2015 edition by the introduction of DL systems, which after the 2019 edition were less used compared to new Transformer-based systems. Our review allowed us to summarize the proposed systems by usage frequency through the 2013-2021 SemEval editions, below is the result of our categorization:

- High usage frequency: LSTM and CNN
- Average usage frequency: BERT and SVM
- Low usage frequency: ALBERT, RoBERTa, XLM-ROBERTa, ERNIE, RNN, GRU, XGBoost, random forest, and logistic regression

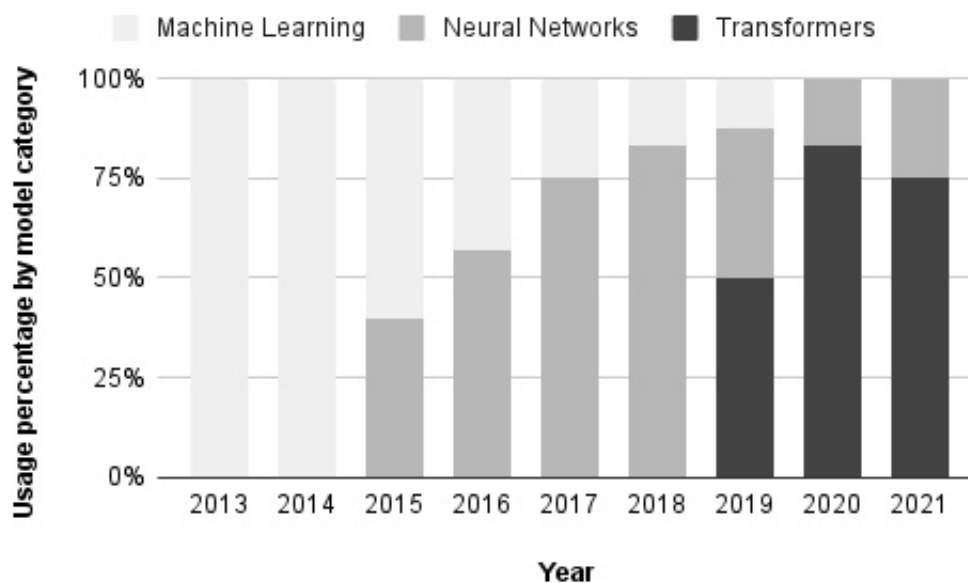


Figure 4. Classification model usage evolution

3.5. Evaluation metrics

The evaluation measures are determining factors in assessing the quality of sentiment analysis models. Model type, data variety, and size advise the evaluation metrics to be used [98]. In the SemEval competitions the F1-score was the most used evaluation metric (in 7 out of 9 studied competitions), other metrics were used including the average recall, the accuracy, and the pearson correlation coefficient. The F1-score use was motivated by the strong imbalance between instances of different classes [13], and where other measurements such as the accuracy fail. Complex metrics are used for model evaluation and analysis which provides a deep understanding of models behavior including the confusion matrix and the receiver operating characteristic (ROC) curve.

4. FINDINGS AND DISCUSSION

SemEval remains an important sentiment analysis competition with new challenges and advanced complexity brought in every edition. On the macro-level, top-ranked systems tend to use universally effective approaches that deliver acceptable results [10], researchers' efforts are then focused on fine-tuning models, features engineering, and datasets. Although the lack of training resources, and constrained systems relying only on the provided resources tend to perform better than unconstrained ones [8], the reasons behind this lay in the additional datasets quality, the data labeling process, and the choice of the features. Moreover, the multilingual aspect of sentiment analysis was also explored noting the lack of models generalization due to the structural differences between languages (i.e., Arabic has abundant use of dialect forms and spelling variants) and the limited improvement using existing translation services [99], which affect models performance and motivate the use of language-specific systems [11].

On the micro-level, there are two approaches for features engineering in sentiment analysis systems: a lexicon-based approach that relies on lexicons to drive the polarity of text, and a machine learning approach that learns a classifying model from annotated text [54]. In the lexicon-based approaches, systems are constrained by the limited lexicon size and require human expertise; features engineered using this approach significantly impact the classifiers [35]. Although a richer feature set improves the classification task [34], the overuse of features may lead to model overfitting [35]. Features based on lexicons have shown to be effective, furthermore, word sense features and word clusters are implemented to improve NLP systems [40]. In [37], word clusters proved to be the most important to their model compared to n-grams and POS which didn't add much improvement. Besides, the experiments in [38] show that brown clusters yield the most considerable impact when compared to syntactic features and word unigrams, the latter leading to model performance degradation. In special cases, syntactic features such as punctuation (i.e., multiple exclamations) [41], and negation patterns can be sentiment-bearing and affect the sentiment perception. Furthermore, Twitter-specific features may encode implicit sentiment and bring significant performance improvement [53] in the context of social media sentiment analysis which contains a lot of implicit sentiment.

In the ML-based approaches, word embeddings were demonstrated to be suitable as features compared to lexicon-based ones [68], and scale well to complex language patterns [95]. Word embeddings, including general-purpose Word2vec or GloVe, are the most used, whereas few systems propose custom embeddings fine-tuned on provided datasets [10]. Moreover, the fine-tuning operation encodes additional semantic information and enriches the word representations. However, shows limited ability with domains containing high types and slang such as social media [36]. The embedding size affects the quality of results, large embeddings provide consistent results [28] while higher dimensions may not improve the performance [100]. Another limitation of word embeddings resides in handling out-of-vocabulary (OOV) words which can be partially overcome using character-level embeddings information [101]. Furthermore, starting from certain available training data, the choice of training models is more critical than the choice of features and word representations [37].

Another important aspect of the studied systems is the classification models whose performance has improved over the editions, this can be explained by the advancement of learning methods and the amount of provided training data [10]. Most used approaches for sentiment detection involve methods from machine learning, computational linguistics, and statistics [97], with a three-scale sentiment classification (positive, neutral, and negative) [37] or two points classification (i.e., positive/negative and subjective/objective). Specific classification categories such as subjectivity and objectivity can be more difficult to perform than the default positive and negative. Furthermore, classification from one scale can be the input features of a classification model for another scale (i.e., Content with high positive/negative confidence is less likely to be objective, thus positive/negative output can be used to determine the objectivity/subjectivity) [55].

Classification models differ in complexity, the most straightforward is rule-based relying on expert pattern detection and rules refinement. Besides, they can measure up to other approaches in performance and provide the flexibility of context-centric approaches such as sentiment toward a topic, the author's feeling, or the general mood [66]. Linear models namely SVM and LR dominated the first editions and performed well compared to other models such as NB, RF, and KNN [37]; nevertheless, RF delivered a state-of-the-art performance [102]. The performance of those models can be further optimized using regularization techniques, this was observed in [97] where L2-regularized LR reimplementation has outperformed the original model. Emerging trends in [36] have manifested in the use of DL techniques, especially CNNs and RNNs, and the use of word embeddings such as GloVe and Word2vec. Moreover, the impact of word embeddings usage on the model's performance was apparent as the initialization of networks with random parameters provides

mediocre results while providing good initialization parameters (i.e., pre-trained word embeddings) boost the model's performance [85]. The massive use of LSTM and CNN implementations using Theano [103] and Keras [10], since their combination with word embeddings provided better performance and proved to be well suited for sentence classification problems. Furthermore, strategies of early and late fusion of word embeddings were explored noting limitations including a lack of features correlation modeling in the late fusion and an intensive required training for the early fusion [68]. The architectures of the proposed models were very basic, simply stacked layers of LSTM, BiLSTM, or CNN enhanced with attention, noise, dropout, and regularization techniques. Furthermore, the fine-tuning process focuses usually on selected hyperparameters such as the layers dimension, the number of layers, the learning rate, and the epochs.

In line with current research directions, recent systems use mainly PLMs such as BERT and its variants. PLMs brought noticeable advancement in performance compared to the DL model; a 10% improvement in F1 score with BERT compared to LSTM was noted in [104]. Furthermore, PLM variants such as ALBERT (light BERT version) can yield comparable results to BERT [105] allowing a benefit of time and computational resources optimization, this can be enhanced using domain adaptation which can improve classification performance [106]. Another studied aspect is ensemble techniques which, depending on the models (voters), may lead to better results compared to individual systems [107]. Furthermore, ensemble techniques based on deep learning approaches profiting from the use of word embeddings and neural networks showed promising results [102]. Moreover, the use of stacking, as an ensemble strategy, demonstrated the robustness of systems [47] improving their accuracy [108]. Votes are decided using an average confidence model score, a straightforward scheme, or other strategies such as soft or hard voting [83]. The benefits of using such techniques come with exceptions where it shows performance degradation; an individual voter may outperform other group models which results in overall performance loss, voters should perform at a similar level (compatibility with the nature of the classification task) to ensure potential gains in performance considering the complexity and the intensive computation that come with a high cost of deployment of such models.

The evolution of used techniques in the proposed systems follows a logical path of research trends, and the transition from ML and DL models to current PLMs contributed to the performance improvement and the maturity of sentiment analysis-based systems to meet industry-grade requirements. While the appeal to transformers and neural networks was motivated by the considerable performance gain and the research context, the proposed systems provide less space for innovation and research efforts delegating this to the provided models and focusing on ready-to-use components. Moreover, the scope of intervention remains very limited for researchers to fine-tune pre-trained word embeddings or models, and basic architecture design which usually leads to minor gains in performance and notable complexity and computational cost. Another point to discuss is thematic sentiment analysis which explored, through SemEval competitions, the identification of some language phenomena such as sarcasm, offense, and humor. In this respect, specialized systems were built to benefit from domain fine-tuning and advanced features engineering for specific domains and languages, thus outperforming general-purpose models. The lack of training data and the difficulty of features engineering remains the main issues related to domain-specific models, thus encouraging the use of general-purpose models for rapid prototyping.

The proposed systems, in most cases, agreed on a similar tasks pipeline. Data cleaning and processing, feature engineering and representation, classification, and visualization, are the principal tasks performed by most participants. Besides, the focus was geared toward particular sub-tasks (i.e preprocessing or classification) rather than the whole pipeline. The optimization process seems difficult to assert since the components are interdependent, poor word representation or the use of random preprocessing techniques may lead to overall performance degradation. Furthermore, raise concerns about the approaches that sentiment analysis systems should consider in the design phase, and urge the need for a common framework for sentiment analysis systems.

5. CONCLUSION

In this paper, we study top-ranked systems in the SemEval competition during the 2013-2021 period. The objectives are to provide a timeline tracking the evolution of used sentiment analysis techniques including data representation, preprocessing, and classification. We focused on the five top-performing systems in each edition, various aspects of these systems were analyzed including data preparation, preprocessing techniques, and classification models. Moreover, we provide a diagnosis and summarization of the challenges, key contributions, and limitations of each system to provide an overview of the sentiment analysis research field.

Notable progress was observed in the data preparation phase, organizers capitalized on the provided datasets from each edition and improved the preparation methodology from relying on manual labeling and expert work, the use of crowdsourcing platforms, to the use of techniques such as data augmentation and distant supervision which enrich datasets to overcome the lack of training datasets limitation noted in the first editions. Moreover, we pointed to the active use of preprocessing techniques in most studied systems with less frequency in transformer-based approaches, due to the specific nature of social media content and the gains in performance and computational complexity. Word embeddings such as GloVe and Word2Vec remain the most used techniques for word representations and features encoding, state-of-the-art systems were provided by coupling word embeddings with neural networks. Furthermore, approaches relying on pre-trained language models provided comparable results with those systems and paved the way for their dominance over neural networks and traditional machine learning algorithms.

We believe our work will help researchers access a variety of experimented approaches for sentiment analysis systems in the context of social media, and allow future participants to provide more innovative solutions. In our future work, we intend to cover other submissions considering aspects such as innovation, and research originality to complement our top-performing filter. Moreover, thematic reviews for features engineering, preprocessing techniques, or classification approaches can be envisaged to provide an in-depth analysis of each sentiment analysis task.

REFERENCES

- [1] "Global social media statistics," DataReportal. <https://datareportal.com/social-media-users> (accessed Jul. 12, 2022).
- [2] U. Farooq, T. P. Dhamala, A. Nongillard, Y. Ouzrout, and M. A. Qadir, "A word sense disambiguation method for feature level sentiment analysis," in *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Dec. 2015, pp. 1–8, doi: 10.1109/SKIMA.2015.7399988.
- [3] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, vol. 131, no. 1, Boston, MA: Springer US, 2012, pp. 415–463.
- [4] J. Yang *et al.*, "A survey of knowledge enhanced pre-trained models," *arXiv:2110.00269*, Oct. 2021.
- [5] K. Hamilton, A. Nayak, B. Božić, and L. Longo, "Is neuro-symbolic AI meeting its promise in natural language processing? a structured review," *arXiv:2202.12205*, Feb. 2022.
- [6] R. Gull, U. Shoaib, S. Rasheed, W. Abid, and B. Zahoor, "Pre processing of Twitter's data for opinion mining in political context," *Procedia Computer Science*, vol. 96, pp. 1560–1570, 2016, doi: 10.1016/j.procs.2016.08.203.
- [7] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson, "SemEval-2013 task 2: sentiment analysis in Twitter," in *2nd Joint Conference on Lexical and Computational Semantics*, 2013, vol. 2, pp. 312–320.
- [8] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov, "SemEval-2014 task 9: sentiment analysis in Twitter," in *8th International Workshop on Semantic Evaluation*, 2014, pp. 73–80, doi: 10.3115/v1/s14-2009.
- [9] S. Rosenthal, S. M. Mohammad, P. Nakov, A. Ritter, S. Kiritchenko, and V. Stoyanov, "SemEval-2015 task 10: sentiment analysis in Twitter," *arXiv: 1912.02387*, Dec. 2019.
- [10] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: sentiment analysis in Twitter," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 1–18, doi: 10.18653/v1/S16-1001.
- [11] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 502–518, doi: 10.18653/v1/S17-2088.
- [12] S. Mohammad, F. B. Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: affect in Tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 1–17, doi: 10.18653/v1/S18-1001.
- [13] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "SemEval-2019 task 6: identifying and categorizing offensive language in social media (OffensEval)," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 75–86, doi: 10.18653/v1/S19-2010.
- [14] M. Zampieri *et al.*, "SemEval-2020 task 12: multilingual offensive language identification in social media (OffensEval 2020)," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, pp. 1425–1447, doi: 10.18653/v1/2020.semeval-1.188.
- [15] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: garbage in, garbage out," *PLOS ONE*, vol. 15, no. 12, Dec. 2020, doi: 10.1371/journal.pone.0243300.
- [16] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, "SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 105–119, doi: 10.18653/v1/2021.semeval-1.9.
- [17] W. Ruch, "The sense of humor: explorations of a personality characteristic," Walter de Gruyter & Co., pp. 1–498, 2010.

- [18] "Twitter," Twitter. <https://twitter.com> (accessed Jul. 12, 2022).
- [19] Aesuli, "SentiWordNet." GitHub. <https://github.com/aesuli/SentiWordNet> (accessed Jul. 12, 2022).
- [20] "Amazon mechanical turk." <https://www.mturk.com/> (accessed Jul. 12, 2022).
- [21] "Confidence to deploy ai with world-class training data," Appen. <https://appen.com/> (accessed Jul. 12, 2022).
- [22] "Prolific quickly find research participants you can trust." <https://www.prolific.co/> (accessed Jul. 12, 2022).
- [23] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems*, vol. 21, no. 4, pp. 315–346, 2003, doi: 10.1145/944012.944013.
- [24] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, vol. 2, pp. 427–431, doi: 10.18653/v1/E17-2068.
- [25] S. Hochreiter and J. Urgan Schmidhuber, "Long short term memory," *Neural Computation*, vol. 9, no. 8, 1997.
- [26] A. Vaswani *et al.*, "Attention is all you need," *arXiv:1706.03762*, Jun. 2017.
- [27] R. Artstein, "Inter-annotator agreement," in *Handbook of Linguistic Annotation*, Dordrecht: Springer Netherlands, 2017, pp. 297–313.
- [28] R. F. Astudillo, S. Amir, W. Ling, B. Martins, M. Silva, and I. Trancoso, "INESC-ID: sentiment analysis without hand-coded features or linguistic resources using embedding subspaces," in *9th International Workshop on Semantic Evaluation, co-located with the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 652–656, 2015, doi: 10.18653/v1/s15-2109.
- [29] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "SOLID: a large-scale semi-supervised dataset for offensive language identification," in *Findings of the Association for Computational Linguistics*, 2021, pp. 915–928, doi: 10.18653/v1/2021.findings-acl.80.
- [30] "Kaggle: your machine learning and data science community." <https://www.kaggle.com/> (accessed Jul. 12, 2022).
- [31] "UCI machine learning repository: data sets." <https://archive.ics.uci.edu/ml/datasets.php> (accessed Jul. 12, 2022).
- [32] "GitHub: where the world builds software," GitHub. <https://github.com/> (accessed Jul. 12, 2022).
- [33] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Cooolll: a deep learning system for twitter sentiment classification," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 208–212, doi: 10.3115/v1/S14-2033.
- [34] N. Kökciyan, A. Celebi, A. Ozgür, and S. Uskudarlı, "BOUNCE: Sentiment classification in twitter using rich feature sets," in *2nd Joint Conference on Lexical and Computational Semantics*, 2013, vol. 2, pp. 554–561.
- [35] T. Günther, J. Vancoppenolle, and R. Johansson, "RTRGO: enhancing the GU-MLT-LT system for sentiment analysis of short messages," in *8th International Workshop on Semantic Evaluation*, pp. 497–502, 2014, doi: 10.3115/v1/s14-2086.
- [36] S. Amir, R. Astudillo, W. Ling, M. J. Silva, and I. Trancoso, "INESC-ID at SemEval-2016 task 4-a: reducing the problem of out-of-embedding words," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 238–242, doi: 10.18653/v1/S16-1036.
- [37] T. Günther and L. Furrer, "GU-MLT-LT: Sentiment analysis of short messages using linguistic features and stochastic gradient descent," in *2nd Joint Conference on Lexical and Computational Semantics*, 2013, vol. 2, pp. 328–332.
- [38] S. Amir, M. B. Almeida, B. Martins, J. Filgueiras, and M. J. Silva, "TUGAS: exploiting unlabelled data for Twitter sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 673–677, doi: 10.3115/v1/S14-2120.
- [39] X. Zhu, S. Kiritchenko, and S. M. Mohammad, "NRC-Canada-2014: recent improvements in the sentiment analysis of Tweets," in *8th International Workshop on Semantic Evaluation, SemEval 2014 - co-located with the 25th International Conference on Computational Linguistics*, 2014, pp. 443–447, doi: 10.3115/v1/s14-2077.
- [40] Y. Miura, S. Sakaki, K. Hattori, and T. Ohkuma, "TeamX: a sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 628–632, doi: 10.3115/v1/S14-2111.
- [41] T. Proisl, P. Greiner, S. Evert, and B. Kabashi, "KLUE: simple and robust methods for polarity classification," in *2nd Joint Conference on Lexical and Computational Semantics*, 2013, vol. 2, pp. 395–401.
- [42] K. Gimpel *et al.*, "Part-of-speech tagging for Twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, vol. 2, pp. 42–47.
- [43] "NLTK:: natural language toolkit." <https://www.nltk.org/> (accessed Jul. 12, 2022).
- [44] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60, doi: 10.3115/v1/P14-5010.
- [45] "Keras: the Python deep learning API." <https://keras.io/> (accessed Jul. 12, 2022).
- [46] G. Angiani *et al.*, "A comparison between preprocessing techniques for sentiment analysis in Twitter," in *CEUR Workshop Proceedings*, 2016, vol. 1748.
- [47] V. Duppada, R. Jain, and S. Hiray, "SeerNet at SemEval-2018 task 1: domain adaptation for affect in Tweets,"





- arXiv:1804.06137*, Apr. 2018.
- [48] B. Liu, *Sentiment analysis and opinion mining*. Cham: Springer International Publishing, 2012.
- [49] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*, Elsevier, 1980, pp. 3–33.
- [50] J. L. Tracy and D. Randles, "Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt," *Emotion Review*, vol. 3, no. 4, pp. 397–405, Oct. 2011, doi: 10.1177/1754073911410747.
- [51] I. Bakker, T. van der Voordt, P. Vink, and J. de Boon, "Pleasure, arousal, dominance: mehrabian and russell revisited," *Current Psychology*, vol. 33, no. 3, pp. 405–421, Sep. 2014, doi: 10.1007/s12144-014-9219-4.
- [52] O. Almatrafi, S. Parack, and B. Chavan, "Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014," in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, Jan. 2015, pp. 1–5, doi: 10.1145/2701126.2701129.
- [53] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: building the state-of-the-art in sentiment analysis of tweets," in *2nd Joint Conference on Lexical and Computational Semantics*, 2013, vol. 2, pp. 321–327.
- [54] H. Hamdan, P. Bellot, and F. Bechet, "Lsislif: feature extraction and label weighting for sentiment analysis in Twitter," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 568–573, doi: 10.18653/v1/S15-2095.
- [55] L. Dong, F. Wei, Y. Yin, M. Zhou, and K. Xu, "Splusplus: a feature-rich two-stage classifier for sentiment analysis of Tweets," in *9th International Workshop on Semantic Evaluation, co-located with the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 515–519, doi: 10.18653/v1/s15-2086.
- [56] B. F. Green, P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, "The general inquirer: a computer approach to content analysis," *American Educational Research Journal*, vol. 4, no. 4, Nov. 1967, doi: 10.2307/1161774.
- [57] L. Deng and J. Wiebe, "MPQA 3.0: an entity/event-level sentiment corpus," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1323–1328, doi: 10.3115/v1/N15-1146.
- [58] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2010, pp. 2200–2204.
- [59] F. A. Nielsen, "A new evaluation of a word list for sentiment analysis in microblogs," *CEUR Workshop Proceedings*, vol. 718, pp. 93–98, 2011.
- [60] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of The 2004 ACM SIGKDD International Conference on Knowledge Discovery And Data Mining*, 2004, doi: 10.1145/1014052.1014073.
- [61] S. M. Mohammad and P. Turney, "NRC word-emotion association lexicon," *Sentiment and Emotion Lexicons*, 2013.
- [62] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford University, 2009.
- [63] S. J. DeRose, "The compass derose guide to emotion words." <http://derose.net/steve/resources/emotionwords/ewords.html> (accessed Jul. 12, 2022).
- [64] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta, "The waCky wide web: a collection of very large linguistically processed web-crawled corpora," *Language Resources and Evaluation*, vol. 43, no. 3, pp. 209–226, 2009, doi: 10.1007/s10579-009-9081-4.
- [65] T. Brants and A. Franz, "Web 1t 5-gram," *Linguistic Data Consortium*, 2006, doi: 10.35111/CQPA-A498.
- [66] H. Reckman *et al.*, "Teragram: Rule-based detection of sentiment phrases using SAS sentiment analysis," *2nd Joint Conference on Lexical and Computational Semantics*, vol. 2, pp. 513–519, 2013.
- [67] J. H. Park, P. Xu, and P. Fung, "PlusEmo2Vec at SemEval-2018 Task 1: Exploiting emotion knowledge from emoji and hashtags," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 264–272, doi: 10.18653/v1/S18-1039.
- [68] M. Rouvier and B. Favre, "SENSEI-LIF at SemEval-2016 Task 4: Polarity embedding fusion for robust sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 202–208, doi: 10.18653/v1/S16-1030.
- [69] Z. S. Harris, "Distributional Structure," in *Papers on Syntax*, Dordrecht: Springer Netherlands, 1981, pp. 3–22.
- [70] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1–4, pp. 43–52, Dec. 2010, doi: 10.1007/s13042-010-0001-0.
- [71] G. Paltoglou and M. Thelwall, "A study of information retrieval weighting schemes for sentiment analysis," in *48th Annual Meeting of the Association for Computational Linguistics*, pp. 1386–1395, 2010.
- [72] P. F. Brown, P. V. DeSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-Based n-gram Models of Natural Language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.
- [73] Z. Li, Z. Xiong, Y. Zhang, C. Liu, and K. Li, "Fast text categorization using concise semantic analysis," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 441–448, 2011, doi: 10.1016/j.patrec.2010.11.001.
- [74] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations*, Jan. 2013.

- [75] J. Pennington, R. Socher, and C. Manning, “GloVe: global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
- [76] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl.a.00051.
- [77] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, “Learning sentiment-specific word embedding for Twitter sentiment classification,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, vol. 1, pp. 1555–1565, doi: 10.3115/v1/P14-1146.
- [78] M. E. Peters *et al.*, “Deep contextualized word representations,” *arXiv:1802.05365*, Feb. 2018.
- [79] Y. Liu *et al.*, “RoBERTa: a robustly optimized BERT pretraining approach,” *arXiv:1802.05365*, Jul. 2019.
- [80] K. Ethayarajh, “How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 55–65, doi: 10.18653/v1/D19-1006.
- [81] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, “Synesketch: an open source library for sentence-based emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 312–325, Jul. 2013, doi: 10.1109/T-AFFC.2013.18.
- [82] A. R. Atmadja and A. Purwarianti, “Comparison on the rule based method and statistical based method on emotion classification for Indonesian Twitter text,” in *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, Nov. 2015, pp. 1–6, doi: 10.1109/ICITSI.2015.7437692.
- [83] S. Xu, H. Liang, and T. Baldwin, “UNIMELB at SemEval-2016 tasks 4A and 4B: an ensemble of neural networks and a Word2Vec based model for sentiment classification,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 183–189, doi: 10.18653/v1/S16-1027.
- [84] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “a convolutional neural network for modelling sentences,” *Indian Journal of Pathology and Microbiology*, vol. 51, no. 3, pp. 370–372, Apr. 2014.
- [85] A. Severyn and A. Moschitti, “UNITN: training deep convolutional neural network for Twitter sentiment classification,” in *9th International Workshop on Semantic Evaluation, co-located with the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 464–469, doi: 10.18653/v1/s15-2079.
- [86] L. Chen, W. Ruan, X. Liu, and J. Lu, “SeqVAT: virtual adversarial training for semi-supervised sequence labeling,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8801–8811, doi: 10.18653/v1/2020.acl-main.777.
- [87] J. Ma *et al.*, “MagicPai at SemEval-2021 task 7: method for detecting and rating humor based on multi-task adversarial training,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 1153–1159, doi: 10.18653/v1/2021.semeval-1.162.
- [88] P. Liashchynskiy and P. Liashchynskiy, “Grid search, random search, genetic algorithm: a big comparison for NAS,” *arXiv:1912.06059*, 2019.
- [89] A. Nikolov and V. Radivchev, “SemEval-2019 task 6: offensive tweet classification with BERT and ensembles,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 691–695, doi: 10.18653/v1/S19-2123.
- [90] W. Dai, T. Yu, Z. Liu, and P. Fung, “Kungfupanda at SemEval-2020 task 12: BERT-based multi-task learning for offensive language detection,” in *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics*, 2020, pp. 2060–2066, doi: 10.18653/v1/2020.semeval-1.272.
- [91] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, vol. 1, pp. 328–339, doi: 10.18653/v1/P18-1031.
- [92] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: a lite BERT for self-supervised learning of language representations,” *arXiv:1909.11942*, Sep. 2019.
- [93] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: generalized autoregressive pretraining for language understanding,” *arXiv:1906.08237*, Jun. 2019.
- [94] D. Faraj and M. Abdullah, “SarcasmDet at SemEval-2021 task 7: detect humor and offensive based on demographic factors using RoBERTa pre-trained model,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 527–533, doi: 10.18653/v1/2021.semeval-1.64.
- [95] A. Pelicon, M. Martinc, and P. Kralj Novak, “Embeddia at SemEval-2019 task 6: detecting hate with neural network and transfer learning approaches,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 604–610, doi: 10.18653/v1/S19-2108.
- [96] J. Zhu, Z. Tian, and S. Kübler, “UM-IU@LING at SemEval-2019 task 6: identifying offensive tweets using BERT and SVMs,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 788–795, doi:





- 10.18653/v1/S19-2138.
- [97] M. Hagen, M. Potthast, M. Büchner, and B. Stein, “Webis: an ensemble for Twitter sentiment detection,” in *9th International Workshop on Semantic Evaluation, co-located with the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 582–589, 2015, doi: 10.18653/v1/s15-2097.
- [98] B. El Haddaoui, R. Chiheb, R. Faizi, and A. El Afia, “Sentiment analysis: a review and framework foundations,” *International Journal of Data Analysis Techniques and Strategies*, vol. 13, no. 4, 2021, doi: 10.1504/IJDATS.2021.120100.
- [99] A. Rozenal and D. Fleischer, “Amobee at SemEval-2017 task 4: deep learning system for sentiment detection on Twitter,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, May 2017, pp. 653–658, doi: 10.18653/v1/S17-2108.
- [100] H. Zhang *et al.*, “MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from Twitter,” in *International Workshop on Semantic Evaluation*, 2019, pp. 683–690, doi: 10.18653/v1/s19-2122.
- [101] C. Baziotis *et al.*, “NTUA-SLP at SemEval-2018 task 1: predicting affective content in Tweets with deep attentive RNNs and transfer learning,” in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 245–255, doi: 10.18653/v1/S18-1037.
- [102] J. Deriu, M. Gonzenbach, F. Uzdilli, A. Lucchi, V. De Luca, and M. Jaggi, “SwissCheese at SemEval-2016 task 4: sentiment classification using an ensemble of convolutional neural networks with distant supervision,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 1124–1128, doi: 10.18653/v1/S16-1173.
- [103] “Theano.” <https://theano-pymc.readthedocs.io/en/latest/> (accessed Jul. 12, 2022).
- [104] P. Liu, W. Li, and L. Zou, “NULI at SemEval-2019 task 6: transfer learning for offensive language detection using bidirectional transformers,” in *International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop*, pp. 87–91, 2019, doi: 10.18653/v1/s19-2011.
- [105] G. Wiedemann, S. M. Yimam, and C. Biemann, “UHH-LT at SemEval-2020 task 12: fine-tuning of pre-trained transformer networks for offensive language detection,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, pp. 1638–1644, doi: 10.18653/v1/2020.semeval-1.213.
- [106] S. Sotudeh *et al.*, “GUIR at SemEval-2020 task 12: domain-tuned contextualized models for offensive language detection,” in *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics*, 2020, pp. 1555–1561, doi: 10.18653/v1/2020.semeval-1.203.
- [107] S. Giorgis, A. Rousas, J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, “aueb.twitter.sentiment at SemEval-2016 task 4: a weighted ensemble of SVMs for Twitter sentiment analysis,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 96–99, doi: 10.18653/v1/S16-1012.
- [108] B. Song, C. Pan, S. Wang, and Z. Luo, “DeepBlueAI at SemEval-2021 task 7: detecting and rating humor and offense with stacking diverse language model-based methods,” in *15th International Workshop on Semantic Evaluation, Proceedings of the Workshop*, 2021, pp. 1130–1134, doi: 10.18653/v1/2021.semeval-1.158.

BIOGRAPHIES OF AUTHORS







Bouselham El Haddaoui     is a Ph.D. student specialized in opinion mining and sentiment analysis, independent IT advisor in digital communication and event management software. He has an engineering degree in software engineering from ENSIAS, Mohamed V University in Rabat Morocco, 2012. He can be contacted at email: bouselham.haddaoui@um5s.net.ma.







Raddouane Chiheb     is a professor of higher education, head of the computer science and decision support Department, Ph.D. in Applied Mathematics from Jean Monnet University of Saint-Étienne, 1998. Specialized Master in Computer Science from Insa in Lyon, France, 2001, and the President of the Moroccan Association for Value Analysis. He can be contacted at email: r.chiheb@um5s.net.ma.



Rdouan Faizi     is a full professor at national school of computer science and systems analysis (ENSIAS). He obtained his Ph.D. in English Language and Literature from Mohammed V Agdal University, 2002. Research areas of interest are linguistique, e-learning, business english skills, TOEIC preparation, scientific communication. He can be contacted at email: r.faizi@um5s.net.ma.



Abdellatif El Afia     is a full professor at national school of computer science and systems analysis, He received his M.Sc. Degrees in Applied Mathematics from University of Sherbrook. He obtained his Phd in 1999 in Operation Research from the University of Sherbrook, Canada. Research areas of interest are mathematical programming (stochastic and deterministic), metaheuristics, recommendation systems and machine learning. He is the coordinator of the artificial intelligence engineering branch (2IA) at ENSIAS. He can be contacted at email: a.elfafia@um5s.net.ma.