Journal of Ambient Intelligence and Humanized Computing (2020) 11:97–117
https://doi.org/10.1007/s12652-018-0862-8

ORIGINAL RESEARCH

Sentiment analysis: a review and comparative analysis over social media

Nikhil Kumar Singh[1] · Deepak Singh Tomar[1] · Arun Kumar Sangaiah[2]

Received: 8 February 2018 / Accepted: 9 May 2018 / Published online: 23 May 2018

Abstract
Sentiment analysis is the computational examination of end user's opinion, attitudes and emotions towards a particular topic or product. Sentiment analysis classifies the message according to their polarity whether it is positive, negative, or neutral. Recently researchers focused on lexical and machine-learning based method for sentiment analysis of social media post. Social media is a micro blogger site in which end users can post their comment in slag language that contains symbols, idioms, misspelled words and sarcastic sentences. Social media data also have curse of dimension problem i.e. high dimension nature of data that required specific pre-processing and feature extraction, which leads to improve classification accuracy. This paper present comprehensive overview of sentiment analysis technique based on recent research and subsequently explores machine learning (SVM, Navies Bayes, Linear Regression and Random Forest) and feature extraction techniques (POS, BOW and HASS tagging) in context of Sentiment analysis over social media data set. Further twitter data-sets are scrutinized and pre-processed with proposed framework,which yield intersecting facts about the capabilities and deficiency of sentiment analysis methods. POS is most suitable feature extraction technique with SVM and Navie Bayes classifier. Whereas Random Forest and linear regression provide the better result with Hass tagging.

Keywords Sentiment analysis · SVM · Navies Bayes · Linear Regression · Random Forest · POS · BOG · HASS tagging

1 Introduction

In the era of Web 2.0, user generated data over the Internet has expanded exponentially. Social Media podium and commercial website, such as Facebook, Instagram, Twitter, LinkedIn, Amazon, Flip cart etc. offer a platform to share their experiences, knowledge and views on recent trend of politics, economics and other global- critical issue (Smailovic et al. 2014; Lau et al. 2014; Li et al. 2014; Rill et al. 2014). Embedding the social intelligence from enormous online comment is a tedious job for any society or individual. These problems lead to develop a social analytic method to automatically extract, analyze, and summarize user-generated data know as Sentiment analysis.

Sentiment analysis (SA) accumulates on-line documents ranging from twitters, face book; product reviews blogs and other social media platform. Online document assist to understand customers attitudes, opinion and emotion. The word sentiment analysis? introduced by Das and Chen (2001) for stock market sentiment analysis. Since then its effect can be grasped in many real-world applications, ranging from studying product reviews (Stepanov and Riccardi 2011) to foreseeing sales and stock markets using social media monitoring (Yu et al. 2013) and analyzing product marketing or political issue (Feldman 2013). Sensitive webpage classification for content advertising (Jin et al. 2007), commonsense based intelligence system interface (Liu et al. 2003), predicting movie sales (Mishne and Glance 2006), Prediction of hostile or negative sources (Abbasi and Chen 2007), E-rule making (Cardie et al. 2006) opinions on a law before its approval, classification of email on the basis of emotion like anger email, depresses email, normal email (Carro et al. 2012) and visual SA for abstraction of

✉ Arun Kumar Sangaiah
arunkumarsangaiah@gmail.com

Nikhil Kumar Singh
nikhilsinghmanit@gmail.com

Deepak Singh Tomar
deepaktomarmanit@gmail.com

[1] Department of Computer Science, Maulana Azad National Institute of Technology, Bhopal, India

[2] School of Computing Science and Engineering, Vellore Institute of Technology, Vellore 632014, India

subjectivity in the human recognition process (Joshi et al. 2011) e.g., objects classification (He et al. 2016), scene recognition (He et al. 2016).

Sentiment extraction from still image and video is more difficult than many other visual recognition tasks. Recently for multimodal sentiment analysis merge visual feature and decision-level fusion methods for affective information extracted from multiple modalities (Poria et al. 2016). Most accessible schemes employ CNNs to extract global sentiment viewpoint from visual (Yang et al. 2018).

This paper gives a review and comparative analysis survey, that beneficial for new comer researchers. This survey provide comprehensive summary of supervised sentiment analysis technique and explores classification (support vector machine and Navies Bayes) and feature extraction techniques (POS, BOW and HASS tagging) in context of Sentiment analysis over social media data set. This paper presents a framework for scrutinize and preprocess, twitter datasets that yield intersecting facts about the capabilities and deficiency of sentiment analysis methods.

The rest of the paper is organized as follows: Sect. 2 presents over view of sentiment analysis; Sect. 3 covers machine learning, lexicon and hybrid based sentiment analysis technique Sect. 4 covers related work on opinion mining, sentiment analysis and polarity detection over modalities like Micro blogger text, multi-lingual and active learning. Section 5 describes the datasets used and proposes an overview of the experiment; next, Sects. 6 and 6.1–6.4 explain how social media data are processed, step for prepossessing, feature extraction technique for efficient SA and experimental Contents for performance evaluation respectively. Section 7 illustrates possible research avenue for sentiment analysis over social media and finally, Sect. 8 concludes the paper and outlines the founding and future work.

2 Sentiment analysis and polarity detection

Sentiment analysis and opinion mining reflect nearly the same meaning. Sentiment analysis identifies emotion expresses in a natural language text and opinion mining is employed to extract the opinion from text. Textual information can be classify into two types first one is factual and second one is opinionated information. Facts are objective sentences and subjective sentences contain explicit opinions, experience, and views about specific product or entities. For example consider two sentences S_1 and S_2.

$$\text{Objective Sentence } (S_1) : \text{“}\textit{I bought an android Phone last week.}\text{''}$$
$$\text{Subjective Sentence } (S_2) : \text{“}\textit{It is such a nice phone}\text{''}$$

where S_1 is an objective sentence which contains the fact about the phone ie *I bought an android Phone*, whereas S_2 is a subjective sentence contains the opinion about that phone ie Phone is very nice? Subjective sentences, classify as positive or negative polarity, for instance sentence S_2 and S_3 contain positive and negative polarity respectively.

$$S_2(+\text{ve}) : \text{“}\textit{It is such a nice phone}\text{''}$$
$$S_3(-\text{ve}) : \text{“}\textit{The battery life was not long}\text{''}$$

Sentiment analysis mostly focuses on analyzing polarity value of subjective sentences. However with some regards subjective sentences dont have any explicit opinions and objective sentence have.

$$\text{Subjective Sentence } (S_4) : \text{“}\textit{I think he came yesterday}\text{''}$$
$$\text{Objective Sentence } (S_5) : \text{“}\textit{My phone broke in the second day}\text{''}$$

For instance S_4 is a subjective sentence but dont have any polarity. Where as S_5 refer to objective sentence that contain implicit negative polarity. This situation lead to extract prime candidates having a semantic orientation that would be leaning more towards objective than to subjective and classify objective sentences according to their sentiment polarity (Appel et al. 2016).

Sentiment analysis has been investigated mainly at three levels document level (Montejo-Rez et al. 2014a; Bravo-Marquez et al. 2014; Ortigosa et al. 2014; da Silva et al. 2014b; Colace et al. 2015; Liang et al. 2014; Cruz et al. 2014), sentence-level (Balahur and Perea-Ortega 2015; Yan et al. 2014; Zhang et al. 2015; Mahyoub et al. 2014; Ptaszynski et al. 2014; Cho et al. 2014), and aspect-level (Kranjc et al. 2015; Lau et al. 2014; Li et al. 2014; Rill et al. 2014; Kang and Park 2014; Wang et al. 2015; Deng et al. 2014; Wu and Tsai 2014; Montejo-Rez et al. 2014b; Lei et al. 2014; Ma et al. 2017). Document level SA is based on one topic-one object assumption i.e. each document concentrated on a single object and contains single user opinions. Its classify the whole document as positive or negative sentiment based on the overall sentiment expressed by opinion holder. For example one person (P_1) posts their comment on their new car that he bought a new car yesterday and having very good experience with it.

Audi Car- User Review (P_1)

> “My Audi CAR was delivered yesterday. It looks fabulous. We went on a long highway drive the very second day of getting the car. It was smooth, comfortable and wonderful drive. Had a wonderful experience with family. Its an awesome car. I am loving it..!”

The person (P_1) presents precise positive opinion about the car, which is suitable for document level sentiment analysis. Document level SA assumes that each document have single entity opinions. So, it is not applicable to documents have multiple entities opinion.

Android Phone—User Review (P_2)

> I bought an android Phone last week. It was such a lovely phone. The touch screen was awesome. The voice quality was pleasant. While the battery backup was not long, that is fine for me. However, my mother was angry with me as I did not consult her before I bought the phone. She also believed the phone was costly, and wanted me to return back to the shop.

For example a person P_2 posts their review about their new android Phone. Here the person P_2 presents descriptive sentiment about android Phone i.e single review may contain multiple opinions even about the same entities. So for fine-grained view of the different opinions expressed in the document about the entities need to move towards sentence level sentiment analysis. Sentence-level SA present sentiment of each sentence of document and try to classify sentence with their polarity value. Before exploring the polarity of sentence, sentence-level SA examines the class of sentence i.e. subjective or objective. Sentence-level SA of P_2 post is shown in Table 1.

Sentiment classification at both document and sentence levels are not sufficient, to understand people likes and/or dislike about the entity. A positive opinion on an object does not mean that the opinion holder likes everything and a negative opinion on an object does not mean that the opinion holder dislikes everything. Document and sentence levels need either the whole document/each individual sentence refers to a single entity. However, in many cases single review of an entity have contained many aspects with different opinion.

Aspect-level SA is a quintuples based sentiment analysis scheme. Use to recognize all quintuples sentiment within a given document. For example in aspect-level SA of P_2 review, start with recognizing all the quintuples as phone, touch screen, voice quality, battery life and price. Then evaluating sentiment polarity of each quintuples shown in Table 2. Recently aspect-level SA is employed by many commercial companies, to identify of all aspects in a corpus of product reviews (Wang et al. 2015; Lei et al. 2014).

Table 2 Aspect-level SA of P_2 post

Quintuples	Polarity
Phone	+ve
Touch screen	+ve
Voice quality	+ve
Battery backup	−ve
Price	−ve

3 Sentiment classification technique

Sentiment classification techniques comes with three different flavor namely machine learning (Balahur and Perea-Ortega 2015; Hogenboom et al. 2014; da Silva et al. 2014b; Colace et al. 2015; Montejo-Rez et al. 2014b), lexicon (Hogenboom et al. 2014; Li et al. 2014; Bravo-Marquez et al. 2014; Kang and Park 2014; Liu and Chen 2015; Balahur et al. 2014; Wang et al. 2015; Wu and Tsai 2014; Liang et al. 2014; Cruz et al. 2014; Ptaszynski et al. 2014; Cho et al. 2014; Lei et al. 2014) and hybrid approach (Yan et al. 2014; Ortigosa et al. 2014; Zhang et al. 2015).

3.1 Machine learning approach

Machine learning approach is an artificial intelligence method that uses to learn computer by either supervised, semi supervised, unsupervised approach or by hybrid approach. Support vector machines (SVM), Naive Bayes, maximum entropy algorithm are come under supervised approach. It preferred to applied when a data and their respective infers statistics about the feature of the data is available. This statistics and their pattern help to make predictions about the future upcoming data. Recently support vector machines (SVM) (Kranjc et al. 2015; Balahur and Perea-Ortega 2015; Yan et al. 2014; Smailovic et al. 2014; da Silva et al. 2014b; Zhang et al. 2015; Wu and Tsai 2014), Naive bayes (da Silva et al. 2014b) and Maximum entropy algorithm (Habernal et al. 2014) has been used along with N-gram (Rong et al. 2014; Balahur and Perea-Ortega 2015;

Table 1 Sentence-level SA of P_2 post

Sentence class	Polarity	Sentence
Objective	N/A	I bought an android Phone last week
Subjective	+ve	It was such a lovely phone
Subjective	+ve	The touch screen was awesome
Subjective	+ve	The voice quality was pleasant
Subjective	−ve	While the battery backup was not long, that is fine for me
Objective	N/A	However, my mother was angry with me as I did not consult her before I bought the phone
Subjective	−ve	She also believed the phone was costly, and wanted me to return back to the shop

Yan et al. 2014), linear regression (Smailovic et al. 2014), Random Forest (da Silva et al. 2014b), and Logistic Regression (da Silva et al. 2014b) to classify sentiment in different platform post like micro blogger post (Das and Chen 2001; Kranjc et al. 2015; Yan et al. 2014; Hogenboom et al. 2014; Montejo-Rez et al. 2014a), bilingual (Jin et al. 2007), multilingual post (Yu et al. 2013; Abbasi and Chen 2007; Cho et al. 2014) stock exchange (Liu et al. 2003) and market analysis post (Mishne and Glance 2006; Balahur and Perea-Ortega 2015; Lau et al. 2014). Semi-supervised (Mishne and Glance 2006; Lei et al. 2014) and unsupervised techniques (Fernndez-Gavilanes et al. 2016; da Silva et al. 2016; Bagheri et al. 2013; Martn-Valdivia et al. 2013) are applied when training set of labeled item are not available to classify the rest of items.Recently researcher focuses to predict the sentiment class of textual information by using unsupervised dependency parsing (Fernndez-Gavilanes et al. 2016), syntax-based rules (Vilares et al. 2017), latent dirichlet allocation(LDA) (Garca-Pablos et al. 2018; Huang et al. 2017; Colace et al. 2015), word embedding and bootstrapping (Garca-Pablos et al. 2018) over state-of the- art of unsupervised methods. Whereas hybrid approaches, combine supervised and unsupervised techniques, or even semi-supervised techniques, to classify sentiments (Balahur and Perea-Ortega 2015; Montejo-Rez et al. 2014b).

3.2 Lexicon based methods

Lexicon based methods is a symbolic technique that based on manually crafted rules and lexica. Lexica are a set of already known idioms, pre-compiled sentence term or phrases used in communication genres. Opinion phrases and idioms together are called opinion lexicon. Collection mechanism of opinion lexicon classifies Lexicon based method into three different mechanism one Manual craft and two automated namely Dictionary and Corpus (ontologies) based approach. Manual approach is very time consuming and May not efficient to used. The ontologies (Mishne and Glance 2006; Smailovic et al. 2014), and dictionaries measuring the semantic orientation of words or phrases (Hogenboom et al. 2014; Bravo-Marquez et al. 2014; Ortigosa et al. 2014; Wang et al. 2015; Zhang et al. 2015; Lei et al. 2014; Jha et al. 2017) mostly used for collecting and gathering opinion lexicon. Recently researchers focus to extract sentiment score form features and customer-review based dictionaries like McDonald financial sentiment dictionary (Li et al. 2014), Harvard IV-4 sentiment dictionary (HVD) (Li et al. 2014), Dalian University of Technology Sentiment Dictionary (Kang and Park 2014; Liu and Chen 2015), National Taiwan University Sentiment Dictionary (Kang and Park 2014; Liu and Chen 2015) and How-Net Dictionary (Liu and Chen 2015) to measure the semantic orientation of words or phrases that can used generating opinion lexicon.

Whereas social media sites like twitter (Kranjc et al. 2015; Balahur and Perea-Ortega 2015; Smailovic et al. 2014; Rill et al. 2014; Colace et al. 2015) and facebook (Ortigosa et al. 2014; Colace et al. 2015) API can be used as web ontology. Apart from that Senti Word Net (Hogenboom et al. 2014; Zhang et al. 2015; Cruz et al. 2014; Ziegelmayer and Schrader 2012), Word Net (Kang and Park 2014; Zhang et al. 2015; Mahyoub et al. 2014) are the lexical database for English and English nouns, verbs, and adjectives are well organized into synonym sets, each representing one underlying lexical concept. Apart from that Francesco Colace (Colace et al. 2015) presents a probabilistic lexicon Sentiment Grabber based on the latent Dirichlet allocation (LDA) (Liang et al. 2014). In that approach LDA used to extract set of documents, graph, the mixed graph of terms (MGTs), belonging to a same knowledge domain. Here MGTs use as structure for the sentiment classification of textual documents. Once reference MGTs have been trained from training documents for a given sentiment orientation, then reference MGT has been used as sentiment filter to classify a new document.

3.3 Hybrid approach

Zhang et al. (2015) proposed a hybrid approach based on word2vec and SVMperf. Initially use word2vec to cluster the similar features in selected domain. And then use lexicon-based and part-of-speech based feature selection methods to generate the training file. Finally classify the comment texts using word2vec and SVMperf. Ortigosa et al. (2014) presents an hybrid sentiment classifier by combining lexicon and machine learning technique. Author developed face book application SentBuk that retrieves massage, comment and like on user profiles. Then classify comment on sentiment polarity (positive, neutral or negative). Ghiassi et al. (2013) present a hybrid Sentiment Analysis approach that use supervised N-gram technique for feature extraction and dynamic artificial neural network (DAN2) algorithm for Twitter-specific lexicon.

Recently SA research is mainly focuses on three different approaches which rely on sentiment lexicons, machine learning and hybrid. But on other hand Caro and Grella (2013) introduces a context-based model for SA, which tuned users' sentiments (or opinions) according to some context of analysis. Context-based model for SA use syntactic-based propagation rules for transferring the sentiment values among the words within the dependency parse tree. Luigi Di Caro present a system called SentiVis. SentiVis implements context-based model for SA that directly leans on data visualization and by transforming the extracted sentiment-knowledge according to a user query. But context-based model enable to handle complex syntactic analysis of conditional users' sentiments (e.g., if it had been good, I would have returned).

4 Background and literature survey

In sentiment classification, computational analysis of people's sentiment, attitudes and emotions about a post need to provided label over the post. Due to dynamic nature of social media data, it's very hard to synchronized change in people's sentiment, attitudes and emotions about a post over time. This section explain and summarized the literature of recent trend in SA with micro bloggers, multilingual and active learning aspect. This section also incorporate recent trend of feature extraction technique of social media data for efficient sentiment analysis.

Along with that total thirty one articles presented in this survey are summarized in Table 3 that contains eleven columns. The main task of the articles is illustrated in the third column. Column fourth illustrates classification level of sentiment analysis. Where "*D*", "*A*" and "*S*" is used to represent document level, aspect level and sentence level respectively. Column fifth and sixth illustrate method and algorithm used for sentiment analysis in different application. Whereas eighth column describes the name of data set and its source that has been used for evaluating different methodology.

4.1 Micro bloggers post sentiment analysis

Micro bloggers post sentiment analysis and opinion mining is a hot research topic. Micro-blogging is a new form of communication that is gaining adherents every day. Millions of people sharing their thoughts everyday on podiums like twitter and face-book and generate billions of messages, which reflect peoples opinions and attitudes (Kranjc et al. 2015; Montejo-Rez et al. 2014a; Bravo-Marquez et al. 2014; Liu and Chen 2015; da Silva et al. 2014b; Wang et al. 2015).

Language variation and short length features of micro blogs generate numerous challenges for sentiment analysis over such noisy data. One challenge is data sparsity, others are open-domain and data dynamics. Data sparsity problem means micro blogs contain a large number of irregular and ill-formed words. Open domain problem focus on domain in-dependency of users post. User can post about any topic not to be restricted to post comment on studied domain only. One more serious problem is data dynamics, as micro blogs post are generated incessantly by a huge and uncontrolled number of users. Data dynamics lead to difficulty to processed and analyzed micro blogging data in real-time.

A Montejo-Rez present semantic based crowd explicit sentiment analysis (CESA) (Montejo-Rez et al. 2014a). In CESA micro blogger texts are scratch using regular expressions. It generates massive stream of micro-blog posts to a textual symbolization of a sentiment with clear polarity value (e.g. annoyance, happens, sadness, confusion, etc.). Then CESA can index new posts by these emotional states on the basis of polarity score of their textual representation.

Kranjc et al. (2015) introduces a cloud based work flow platform for dynamic adaptive on-line sentiment analysis of micro bloggers post. This workflow minimizes the effort required for tweets labeling and provides an easy way to share the results. Work flow platform is able to handle change in data stream and adapt it component over time. Workflow platform manage the dynamic nature of sentiment classifier by continuous update through active learning and support vector machine. Work flow platform support bi polar polarity for sentiment classification ie either positive or negative. But tweets can also be neutral (Kranjc et al. 2015).

For micro-blogger tweets, sentiment analyses work on sentence level. Where every single tweet can be consider as static statement and assigning a polarity score to the entire tweet. Whereas Kontopoulos et al. (2013) proposed aspect level SA for micro-blogger tweets and use ontology for evaluating the aspects of the tweets. And breaking down each tweet into a set of aspects relevant to the subject.

4.2 Multi-lingual sentiment analysis

On social media platform, peoples who share their comment and opinion belongs from different backgrounds and from different cultures. And use their own language to share their opinion which leads multilingual opinion mining systems. Recently some bilingual and multilingual (Balahur and Perea-Ortega 2015; Hogenboom et al. 2014; Cruz et al. 2014; Balahur and Turchi 2014; Xiao et al. 2017) sentiment analysis scheme is developed. For bilingual SA (Yan et al. 2014) two type sentiment classifier is available monolingual and multilingual. In monolingual sentiment classifier, sentiment lexicon is available only for one reference language and other target language is to be translated into reference language. Multilingual sentiment classifier incorporate the multilingual data and acquired a multilingual classifier that achieved enhanced the classification rate than the monolingual classifiers taken independently (Balahur and Perea-Ortega 2015). Yan et al. (2014) builds machine learning based bilingual approach to conduct sentiment analysis on both Chinese and English tweets. Instead of processing English and Chinese comments individually. Bilingual approach treats tweets as stream of text containing both Chinese and English words. This bilingual approach analyzing movie reviews twitter as stems of words and generate feature vectors after preprocessing. Apply two exchangeable natural language models, SVM and N-Gram to classify tweets.

Table 3 Article summery

R	T	L	M	Algo	P	Data set	Lang	A	Future scope
Zhang et al. (2018)	Micro blog SA	D^{*}	LA^{+}	DBA	SM	Micro blogger text	Chinese	74.8%	Topic predictions
Li et al. (2018)	Tourism SA	A^{***}	LA	PMI, assembled mutual information	SM	Tourism review	Chinese	86.50%	Domain Specific Sentiment Lexicon
Garca-Pablos et al. (2018)	Multilingual SA	A	ML^{++}	LDA, bootstrapping	SM	Restaurant and hotel review	English, Spanish and French	80.5%	Handle negation expression
Huang et al. (2017)	Micro blog SA	D	LA	Latent Dirichlet allocation	–	Labeled tweets	Chinese	70.75%	Multilingual SA
Zhou et al. (2017)	Customer preference mining	S^{**}	HA^{+++}	Lexical and rough set theory	–	OMD,HCR	English	–	Ambiguous preference
Vilares et al. (2017)	Multilingual SA	S	ML	POS and Dependency parsing	SM	Taboada and Grieve corpus	English, German and Spanish	75.75%	Universal aspect extraction
Fernndez-Gavilanes et al. (2016)	Customer review	S	ML	Dependency parsing and dictionary based approach	SM	Cornell movie review, Obama-McCain Debate and SemEval-2015	English	69.95%	Ambiguous preference
Appel et al. (2016)	Fuzzy polarity score	S	HA	Naives Bayes, Maximum entropy	–	Twitter data set, movie review	English	88.02%	Work over humors, metaphors, similes, sarcasm and irony
Cui et al. (2016)	Distributed learning SA	S	ML	BOG, SVM, POS	–	Chinese financial news	English	93.18%	Weighted feature vector
Khan et al. (2016a)	SA and polarity classification	A	ML	SWN, POS, SVM	SM	Cornell MRD,MDSD	English	80.94%	Classify tweets also as neutral
Khan et al. (2016b)	SA Polarity Detection	S	ML	SWN,PMI,POS	SM	Cornell movie review	English	84%	Sentiment domain specific word
Kranjc et al. (2015)	Dynamic SA	A	ML	SVM	Cloud	TD	English	76.04%	Classify neutral tweets
Rong et al. (2014)	Auto-encoder SA	A	ML	Bagging, Skip Gram	SM	IMDB	English	87.73%	Initialize prediction weights
Balahur and Perea-Ortega (2015)	Multilingual SA	S	HA	SVM, ngram, dba	SM	SemEval, TASS 2013	English, Spanish	64.23%	Slag dictionaries for non base language
Yan et al. (2014)	Bilingual SA	S	HA	SVM, N-gram, Dba	SM	Movie review	English, Chinese	98%(E), 85% (C)	Expression differences over culture
Smailovic et al. (2014)	Stock prices prediction	A	ML	SVM, LR	SM	Baidu, Stanford	English	F-measure 0.671	Micro twitters in financial domain
Lau et al. (2014)	Extraction of market intelligence	A	LA	Fuzzy, ontology mining	SM	Consumer review, product descriptions	English	79.10%	Credibility and quality of customer comments
Hogenboom et al. (2014)	Multilangual SA	S	HA	POS Tagger, SentiWordNet	SM	Movie review	Dutch, English	62.2%	Cross lingual sentiment propagation process

Table 3 (continued)

R	T	L	M	Algo	P	Data set	Lang	A	Future scope
Montejo-Rez et al. (2014a)	Crowd explicit SA	D	LA	Gain ratio, cosine distances	SM	Emoticon, SFU review	English, Spanish	71.86, 69.75%	Polarity calculation of emoticon
Li et al. (2014)	SA in financial domain	A	LA	Dictionary based approach	SM	News, stock quotes	English	48%	Predicting time horizon for financial domain
Bravo-Marquez et al. (2014)	Micro blogger post SA	D	LA	POS	SM	Stanford, Sanders, SemEval	English	85%	Aspect level SA over micro blogger post
Rill et al. (2014)	Detection of emerging political topics	A	LA	Hashtags, binomial distribution	SM	Twitter Data, Google Trends	German	68%	Tweets geo-information
Kang and Park (2014)	Customer satisfaction	A	LA	Dictionaries based approach	SM	App Store HQ	English	–	Customer requirement
Liu and Chen (2015)	Multi-Label SA for Micro Blogger	D	LA	Dictionaries	–	Hr, Ia	Chinese	52.3%	Multi-label sentiment analysis for micro blogger
Ortigosa et al. (2014)	Student SA	D	HA	POS, Euclidean distance	SM	Facebook API	Spanish	83.27%	Change towards negative sentiment
da Silva et al. (2014b)	Micro tweet SA	D	ML	Naive Bayes, SVM, Random Forest, and Logistic Regression	Weka	Twitter Corpus, OMD, HCR	English	87.20%	Classify tweets also as neutral
Wang et al. (2015)	Random subspace SA	A	LA	POS	Weka	Sentiment analysis dataset	English	85.55%	Parallel computing
Deng et al. (2014)	Supervised Weighted SA	A	ML	Term frequency, BoG	–	Movie review, product review	English	88.85%	Stop words removal
Colace et al. (2015)	Graph based SA	D	HA	LDA, Mixed graph of terms	Python	Movie review	English	88.50%	Synonyms from annotated lexicon
Zhang et al. (2015)	Chinese SA	S	HA	SVM, POS, Word2vec	SM	Product review	Chinese	90%	Structured information of sentence
Wu and Tsai (2014)	Concept net-based SA	A	LA	Dictionaries, SVM	SM	Human intelligence tasks	English	79.3%	Standard deviation of sentiment values in the sentiment dictionary
Mahyoub et al. (2014)	Arabic SA	S	ML	SVM, NB, arabic word net	Python, rapid miner	Movie review, book review	Arabic	97%	Dialect and special regional words
Ptaszynski et al. (2014)	Japanese SA	S	LA	Dictionary, ML-ASK, CAO	Japanese corpus	YACIS, Nakamuras dictionary	Japanese	86%	Classify tweets also as neutral
Cho et al. (2014)	Multiple Dictionaries based SA	S	LA	Dictionary and data driven	Standford core NLP suit	Smart phone, movie, product review	English	82.6%	Data driven sentiment dictionary
Montejo-Rez et al. (2014b)	Rank graph based SA	A	HA	Senti word net, Random Walk, SVM	SM	Twitter API	English	F-measure 0.6285	Automatic spell checker

Table 3 (continued)

* Document level sentiment analysis; ** Sentence level sentiment analysis; *** Aspect level sentiment analysis;
+ Lexicon based approach; ++ Machine learning approach; +++ hybrid approach

Balahur and Perea-Ortega (2015) develops multilingual sentiment analysis system for English and Spanish language. Alexandra adapts English comment to Spanish, by employing in-house built dictionaries and machine-translated data for training. And combining multilingual data and obtained a multilingual classifier. Performance of multilingual classifier is overall better than the monolingual classifiers taken separately.

Higher linguistic processing (lemmatization, stop word removal) actually deteriorates the performance. Yu et al. (2013) require minimal linguistic processing and use unigrams and bigrams to analyze revisions in the polarity of the sentiment as negation positive and intensifier negation.

Hogenboom et al. (2014) presents semantics-guided cross-lingual sentiment mapping approach. In this work author explore that sentiment tends is partly language specific, and recently research objective is to explore comparability of sentiment scores across language. Hogenboom et al. (2014) chose one reference language for which sentiment lexicon is available. Translate target language into reference language by using sentiment lexicon that prepared by propagating sentiment of seed words in a semantic lexicon for the target language. Then subsequently analyze the translated text by mapping sentiment scores from a semantically enabled sentiment lexicon available for reference language.

4.3 Active learning for sentiment analysis

In the era of Web 2.0, relationship between public sentiment and e- commerce is analyzed using the consumer comments at social media or electronic commerce Web sites. Consumer comment towards a product and companies use to predict future design strategies and stock price changes. Recently number of researcher adopted active learning approach to predict market strategy by analyzes the sentiment score of tweet streams at social media or electronic commerce Web sites.

Smailovic et al. (2014) present a stream based active learning approach to predict stock price changes by analyses sentiments in stock related tweets. This active learning based on Granger causality test that state sentiments in stock related tweets can be used as indicators of stock price movements a few days in advance. In this approach SVM classifier is used to categories twitter posts into three sentiment categories of positive, negative and neutral.

Lau et al. (2014) present a social analytics based semi-supervised fuzzy product ontology mining algorithm. Perform a fine-grained extraction of market intelligence to improve product design and marketing strategies.

Li et al. (2014) present an lexicon based approach to analyze the news impact from sentiment dimensions by generating an generic stock price prediction framework. Generic framework use Harvard psychological dictionary and LoughranMcDonald financial sentiment dictionary to construct the sentiment dimensions.

Ortigosa et al. (2014) present a users emotional state based adaptive e-learning systems by using hybrid approach. Extract students sentiments towards a course that can be use as feedback for teachers, especially in the case of online learning for adaptive systems.

Nassirtoussi et al. (2015) work on the challenges to predicts fundamental hidden relationship between news and the stock exchange. Nassirtoussi proposed a system that bringing together natural language processing and statistical pattern recognition as well as sentiment analysis to predict directional-movement of a currency-pair in the exchange market based on the words used in adjacent news-headlines in the previous few hours.

4.4 Feature extraction technique

Recently for polarity detection of text many machine learning approach such as SVM, probabilistic model has been proposed. But because of curse of dimension ie high dimension nature of text, there still a research gap that inspired dimensionality reduction and feature extraction. An intuitive idea is to extract the features of each text instead of whole massage having two steps. First is to elect feature sets and then extract feature values. There are several methods to elect representative word sets, including Chi square (Liang et al. 2014), local/global document frequency, bag-of-words (Rong et al. 2014; Balahur and Perea-Ortega 2015; Yan et al. 2014), feature hashing (da Silva et al. 2014a; Rill et al. 2014) and information gain (Habernal et al. 2014).

Rong et al. (2014) proposed AEBPA architecture employs Skip-gram auto-encoder for dimensionality reduction and unsupervised approach for feature learning. Some challenges still need to be addressed i.e. better initialize the weights of the model. And employ more data to train a better embedding to use more sentimental corpus for the experiment. Habernal et al. (2014) explores and evaluates various preprocessing, features extraction and classifier technique. And present a hybrid method that achieved an F-measure of 0.69 using a combination of features (unigrams, bigrams, POS features, emoticons, character n-grams) and preprocessing techniques (unsupervised stemming and phonetic transcription). Liu and Chen (2015) presents a multi-label classification based approach for sentiment analysis. The prototype system has three main components, text segmentation, feature extraction, and multi-label classification. The features extraction included raw segmented words and sentiment features based on the three different sentiment dictionaries, Dalian University of Technology Sentiment Dictionary, National Taiwan University Sentiment Dictionary and HowNet Dictionary, and the bag of words is the feature representation. Wang et al. (2015) proposed a random subspace method (POS-RS) for sentiment analysis based on part-of-speech analysis. POS-RS maintain the balance between the accuracy and the diversity of base learners. By introducing two important parameters content lexicon subspace rate and function lexicon subspace rate. POS-RS can reduce bias and variance simultaneously.

5 Data set

The data sets used in SA are important issues in these field. The main sources of data are from the product reviews as show in Tables 3 and 2. Tables 3 and 2 contains detail about variety of data set that has been used in different application. Description of data set mention in Table 3 is describe with literature survey section. Data set Table 2 that contains six columns. The size and polarity length are illustrated from third to six column. Size attribute represent number of comment and average length reflect their word count available for sentiment analysis in respective data set.

The main sources of product review are social networking site. That provided their API application like twitter API and facebook API to fetch people opinion from social network. These product reviews are essential for corporate and industrial sector to take proper commercial decisions about their products. SA is not limited only to product reviews but also be helpful for analyzes stock market influence on economy (Smailovic et al. 2014; Lau et al. 2014), news articles (Kang and Park 2014; Lei et al. 2014), and influence of political issue in elections result (Rill et al. 2014).

Table 4 Data set description

Ref.	Name	Size/bumber	Average, length	+Ve	−Ve
1	Cornell Data set, Version 1.0	1400	655	700	700
2	Cornell Data set, Version 2.0	2000	656	1000	1000
3	English Movie Reviews	1000	354	689	308
4	Reviews of Hindi Movies	1000	323	741	254
5	Blog Posts on Libyan Revolution	1486	1130	551	680
54	Stanford Dataset	1,600,000	1171	800,000	800,000
55	Twitter Sentiment Corpus Data set	1224	44	570	654

In political debates (Rill et al. 2014) it is seen that many social networking site like face-book and twitter present a great impact. Particularly, offers a platform where debates on inflame topics can be identified prior than other standard information channels. Social media and blog are considered to be rich source of information where people free to share their opinion and view about a certain topic.

6 Comparative analysis

Comparative analysis are present interesting and useful facts regarding the state-of-the-art of sentiment analysis. For comparing the all the basic stand-alone classifiers such as SVM, Naives Bayes, Linear Regression and Random Forest this paper present an feature extraction and preprocessing framework. This framework preprocessed the Stanford dataset (Go et al. 2009) has 1,600,000 training tweets collected by a scraper via Twitter API and apply bag-of-words (BoW), feature hashing (FH), and POS feature extraction technique evaluate the potential of ensembles and boost classification accuracy.

6.1 Preprocessing of social media data

Social Media community has its own specific slag language to post massage where massage contains symbols, misspelled words and sarcastic sentences. Therefore pre processing of social media specific data is important in sentiment analysis (Haddi et al. 2013). Classification accuracy can be improved with appropriate selection of preprocessing techniques (Balahur and Perea-Ortega 2015; Smailovic et al. 2014). This paper explored the unique properties of slag language and experimented with the user and topic replacement, word normalization, web link replacement, stop word removal, slag replacement and negation to better define the feature space (Fig. 1).

(1) User and topic labeling: User and topic name don't have any sentiment value. In order to produce quality data, the users cited in the tweet and marked with symbol "@ ", are replaced with "*PERSON* " and the topics of the tweet, marked with symbol(marked with "#") are replaced with "*TOPIC* ". Consider the comment C_1 and C_2 where username and topic marked with "@″ and

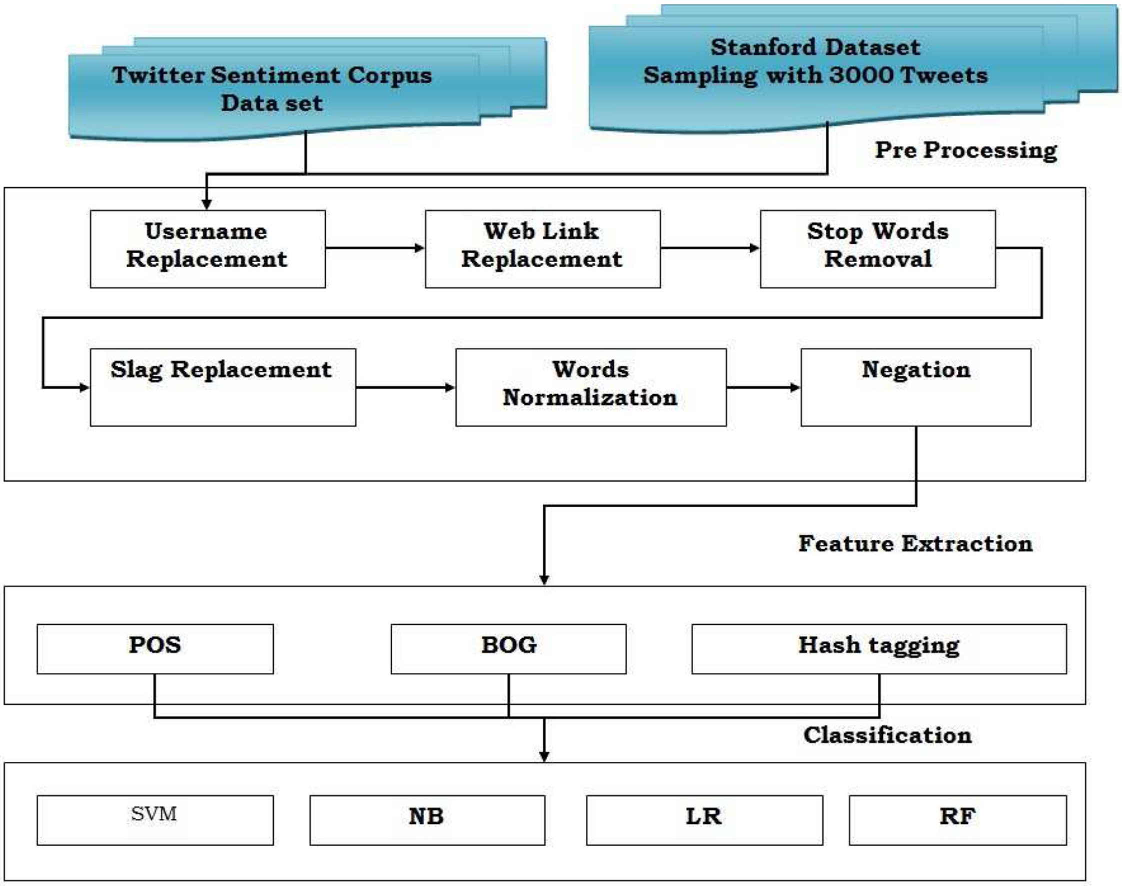


Fig. 1 Feature extraction and prepossessing

"*TOPIC*" are respectively replace by "*USERNAME* " and "*TOPIC*".

Case 1: **Users Mentioned in the Tweet** Comment marked with symbol "@ " are replaced with "*USERNAME* "

Unprocessed Comment (C_1) *-It is such a nice phone "@" Nikhil*

Processed Comment (C_1) *-It is such a nice phone USERNAME*

Case 2: **Topics Mentioned in the Tweet** Comment marked with symbol "#" are replaced with "*TOPIC* ".

Unprocessed Comment (C_2) *-It is such a nice phone "#" Apple*

Processed Comment (C_2) *- It is such a nice phone TOPIC*

(2) Word Normalization (WN): In word normalization, the phase of words are matched with entries in Rogets Thesaurus. If its not matched, recurring letters are sub sequentially compact until its not matched.(e.g. "*perrrrrrrrrrrrrrrrrrfeeect* " becomes "*perrrfeect* ", "*perrfect* ", and subsequently "*perfect*"). Consider the unprocessed comment C_3, C_4 and C_5 where each tokens are compared to entries in Rogets thesaurus and return processed comment C_3, C_4 and C_5.

Unprocessed Comment (C_3): *Perrrrrrrrrrfeeect phone.*

Processed Comment (C_3): *Perfect phone.*

Unprocessed Comment (C_4): *I looooove you.*

Processed Comment (C_4): *I love you.*

Unprocessed Comment (C_5): *Awesome phone, I am Lovvvvvvving it.*

Processed Comment (C_5): *Awesome phone, I am Loving it.*

(3) Slang replacement (SR): Slag replacement over social media include the frequently used semantics in order to normalize the tweet language. For including the frequently used semantics over Social Media, we need to extract list of slang expressions over the globe. Slag replacement is specifically applicable for SMS texts, where words in their original form has little to do with language employed in ordinary texts.Consider the unprocessed comment C_6 where tokens "*D* " are compared to entries in slag meta data and return processed comment C_6 with token "*The*".

Unprocessed Comment (C_6): *D battery life was not long.*

Processed Comment (C_6): *The battery life was not long.*

(4) Stop Word Removal: Stop-words are the words which do not indicate any sentiment, for example "*With*", "*a*", "*There*", "*They*" etc. Since they don't have sentiment value so these are removed in order to produce quality data.

(5) Negations: This paper treats all negation expressions in the same way and replace the negation words (not, isnt, arent, wasnt, werent, hasnt, havent, hadnt, doesnt, dont, didnt) with a unique token "*NEGATION*". Consider the unprocessed comment C_7 where negation word not is replace with "*NEGATION*" and return processed comment C_7. *Unprocessed Comment* (C_7) *(−ve)*: *The battery life was not long. Processed Comment* (C_7) ***(−ve)***: *The battery life was NEGATION long*

6.2 Feature extraction from social media data

Once the massages are preprocessed, processed massages are passed for sentiment classification. For relevant classification this paper deploys bag-of-words (BoW), feature hashing (FH), and POS feature extraction technique to extract and select text features.

(1) Parts of Speech (POS) tagger: POS taggers provide lowest level of syntactic analysis for parsing and word sense disambiguation. Its annotated each word in tweets with a grammatical marker as noun, verb, adjective, adverb, coordinating conjunction etc. In SA adjectives are fine indicators of sentiment. Polarity of adjectives decides the subjective or objective status of tweets and it has been used to guide feature selection for sentiment classification. Consider the unprocessed comment C_8 and their resultant pos tagger provided by Stanford Parser (http://nlp.stanford.edu:8080/parser/index.jsp). In processed comment C_8 word "*nice*" is adjective that shown polarity of comment C_8 about the entity "*phone*".

Unprocessed Comment (C_8): *It is such a nice phone*

Processed Comment (C_8): *it/ **PRP** is/ **VBZ** such/ **PDT** a/ **DT** nice/ **JJ** phone/***NN**

(2) Bag-of-Words (BoW): BoW treats each word (token) in a tweet as order-invariant collection of features and list corpus (C) of tweet words for dictionary (D) as shown in Eq. 1.

$$C = W_1, W_2, W_3, \ldots, W_n \tag{1}$$

Bag-of-word takes unigram, bigram or n-gram phase of words into account and provides specific subjectivity score with help of sentiment lexicon. In sentiment analysis bigram, trigram and n-gram capture better sentiment than unigram. Consider the comment C_9 and focus over word "*highly*", "*recommend*" and "*book*". Unigram work over single word token "*recommend*" whereas bigram take two word phase token i.e. " *highly recommend*" for calculating the polarity of comment

Comment (C_9): *I highly recommend this book.*

Word phase " *highly recommend*" defiantly has high polarity value than "*recommend*"

(3) Feature hashing (FH): Hash tagged words are labels of sentiments and emotions embedded by writer itself. Consider the comment C_{10} where writer itself embedded their sentiment in tweet.

Comment *(C_{10}): Cant wait to have my own Google glasses* ***# awesome***

Hash tag are extremely useful to extract emotion from sarcastic and objective tweets. For example consider the sarcastic comment C_{11} where review is not positive but writer enjoy that moment and in comment C_{12} writer show their anger with hashtag over objective tweet.

Comment (C_{11}): The reviewers want me to re-annotate the data. #**joy**

Comment (C_{12}): Mika used my photo on tumblr. # **anger**

6.3 Sentiment classification of social media data

After extracting relevant feature from social media data processed dataset are passed for sentiment classification. Classifiers SVM, Naives Bayes, Linear regression and random Forest applied separately over different combinations of bag-of-words (BoW), feature hashing (FH), and POS.

(1) Support Vector Machine: Support vector machine is binary classification method that classifies the text data set (td) according to their sentiment into positive and negative classes. The Support vector machine determines the optimal separating hyper plane (WSV + b) between positive and negative sentiment to maximize the margin (m) of the training data as shown in Fig. 2. Text data set (td) is the set of n couple of element (W_i, S_{s_i}),where W_i is associated with word within the text and S_{s_i}) indicate their respective sentiment class (+ve, −ve) as shown in Eq. 2. W_i can be capture by using feature extraction technique such that N-gram, Part of speech and Hass Tag.

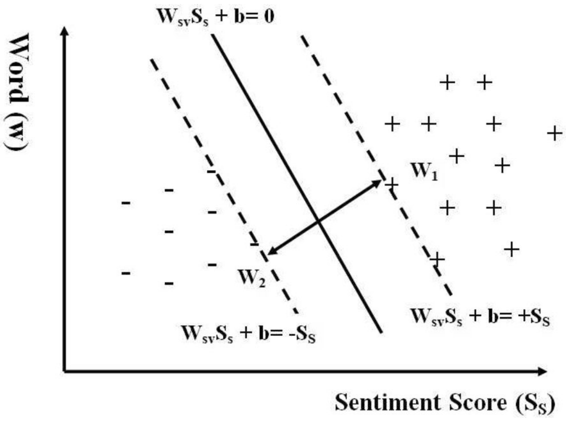


Fig. 2 Sentiment classification using SVM

$$td = (W_i, S_{s_i})W_i \in sv, S_{s_i} \in \{+ve, -ve\}_{i=1}^{n} \tag{2}$$

The hyper plane are define through support sentiment vectors as shown in Eq. 3.

$$\begin{aligned} SentimentVectorspace(sv) = \{&(Good, +ve), \\ &(Nice, +ve), \\ &(Bad, -ve)\} \end{aligned} \tag{3}$$

$$\begin{aligned} d = &\left\{(W_{sv} * S_{s_i} + b_1) - \left(+S_{s_i}\right)\right\} \\ &- \left\{(W_{sv} * S_{s_i} + b_2) - \left(-S_{s_i}\right)\right\} \end{aligned} \tag{4}$$

$$d = \frac{\mid b_1 - b_2 \mid}{\parallel W \parallel} \tag{5}$$

In SVM for optimal hyper plane it is needed to maximize the width of the margin (w). Where Eqs. 4 and 5 shown the positive and negative hyper plane as A and B. if any word W_i belong to sentiment Vector Space have positive S_{s_i} then its reside above hyper plane A and if negative then B. where distance between hyber plane A and B depend upon b_1, b_2 and $\|W\|$ whereas to maximized the margin (d), its needed to minimized weight $\|W\|$ of sentiment vector space (WSV) as shown in Eqs. 6, 7.

$$(W_{sv}.S_{s_i} + b_1) \geq + S_{s_i} \forall W_i \in +S_{s_i}^{A} \tag{6}$$

$$((W_{sv}.S_{s_i} + b_2) \leq - S_{s_i} \forall W_i \in -S_{s_i}^{B} \tag{7}$$

(2) Naive Bayes: Naive Bayes is a probabilistic classifier, meaning that for a text (t), out of all sentiment s ∈ (positive, negative, neutral). Navie Bayes returns the Sentiment SE which has the maximum posterior probability given the x as shown in Eq. 8. Then Bayes rule breaks down the conditional probability P(s | t) into three other probabilities as shown in Eq. 9:

$$S_E = argmax_{S \in (Positive, Negative, Neutral)} P(s|t) \tag{8}$$

$$P(s|t) = \frac{P(t|s)P(s)}{P(t)} \tag{9}$$

where P(s|t) is desire probability, compute the probability of s, the sentiment, given to text t. P (t| s) probability of t, the text, given to sentiment s, determined by previously gathered information. P(s) is independent probability of s (sentiment) and P (t) is independent probability of t (text). Then by substitute the rule breaks down the conditional probability P(s | t) (Eq. 8) into Eq. 8 desire sentiment S_E explore as Eq. 10.

$$S_E = argmax_{S\in(Positive,Negative,Neutral)} \frac{P(t|s)P(s)}{P(t)} \tag{10}$$

P(s|t) will be computed for each sentiment class (positive, negative, neutral) over the same text (t) with the same probability of text P(t). Thus probable sentiment SE for text t is the highest product of prior probability of the sentiment P(s) and the likelihood of the text P(t|s) as shown in Eq. 11.

$$S_E = argmax_{S\in(Positive,Negative,Neutral)} P(t|s)P(s) \tag{11}$$

Availability of training set for text (t) that classified desired text into their specific sentiment categories are relatively low and its not possible to find a specific tweet in the training set always. So in order to calculate P(t|s), tokenize the text and calculate the probability for each word in the training set as shown in Eq. 12.

$$S_E = argmax_{S\in(Positive,Negative,Neutral)} P(w_1, w_2, w_3, \dots, w_n|s)P(s) \tag{12}$$

$$S_E = argmax_{S\in(Positive,Negative,Neutral)} P(w_1, w_2, w_3, \dots, w_n|s)P(s) \tag{13}$$

Regrettably its not practical to compute positioning probability of every possible combination of words (Token) that would require enormous parameters and unbearable large training data sets. In order to overcome this limitation Navie Bayes only encode word identity and not position. The word "*w*" independently has the same effect on classification whether it occurs as the 1st, 2nd, or last word in the text. Text probabilities P($w_1, w_2, w_3 \dots w_n|s$) are independent given the sentiment class (S_E) therefore "*naively*″ multiplied as shown in Eq. 13.

$$P(w_1, w_2, w_3, \dots, w_n|s) = P(w_1|s) * P(w_2|s) * P(w_3|s) * . * P(w_n|s) \tag{14}$$

Then by substitute the rule breaks down the text probability P ($w_1, w_2, w_3 \dots w_n|s$) (Eq. 13) into Eq. 12 desire sentiment *S E* explore as Eq. 14. Beside the classification, sentiment classification of text using naive bayes need to consider word positions, by simply walking an index through every word position in the text as shown in Eq. 15. To avoid underflow and increase computational speed apply log space equation express as Eq. 16.

$$S_E = argmax_{S\in(Positive,Negative,Neutral)} P(s) \prod_{w\in W} P(w|s) \tag{15}$$

$$S_E = argmax_{S\in(Positive,Negative,Neutral)} P(s) \prod_{i\in Position} P(w_i|s) \tag{16}$$

$$S_E = argmax_{S\in(Positive,Negative,Neutral)} logP(s) + P(s) \prod_{i\in Position} P(w_i|s) \tag{17}$$

By considering features in log space Eq. 16 predict the desire sentiment polarity as a linear function of input words.

6.4 Experimental setup

In order to evaluate the performance of the breach mark approach, two experimental campaigns have been carried out. The first one has been performed by using a Stanford dataset (Go et al. 2009) has 1,600,000 training tweets collected by a scraper via Twitter API. Scraper simultaneously sends separate query for both positive and negative emotion. After applying preprocessing, one gets 800,000 tweets with positive and 800,000 tweets with negative emotions. Second experiment has been carried out over Sanders Twitter Sentiment Corpus (Ziegelmayer and Schrader 2012). Twitter corpus is collection of tweet related to four search item @apple, #google, #microsoft, and #twitter. Each tweet has positive, neutral, negative, and irrelevant sentiment label. After applying preprocessing, data set 570 positive and 654 negative tweets. The evaluation measures used in this research are accuracy and improvement factor. Accuracy is proportional to correctly classified tweets as shown in equation 18.

$$Accuracy = \frac{Postive_{Hit} + Negative_{Hit} + Neutral_{Hit}}{T_{Tweets}} \tag{18}$$

where *T tweets* is the total number of tweets used for this experiment and *Positive Hit*, *Negative Hit* and *Neutral Hit* represent total number of Positive, negative and neutral tweets that has been correctly classified.Whereas improvement factor is parentage of improvement with baseline classifier.

Compared results of stand-alone classifiers and different combinations of bag-of-words (BoW), feature hashing (FH), and PoS are described in Table 3. Performance of standalone classifier is to be increase, if it's collaborated with feature extraction technique. Performance evaluation is categorized into two categories and four sub class according to data set and combination feature extraction with classifier ie Standalone Classifier, classifier with POS, Classifier with BOW and Classifier with HT over TSCDS and SFD.

In standalone class Nave Bayes, SVM, Random Forest and Linear Regression gain 53.77, 49.71, 54.54 and 55.12TSCDS and 48.45, 45.23, 51.37 and 52.23 data set

Table 5 Comparative analysis of sentiment analysis technique

Technique	Accuracy	
	Twitter sentiment corpus data set	Stanford dataset
Nave Bayes	53.77	48.45
SVM	49.71	45.23
Random Forest	54.51	51.37
Linear Regression	55.12	52.23
Nave Bayes + POS	83.13	80.12
SVM + POS	83.27	81.34
Random Forest + POS	81.15	78.34
Linear Regression + POS	79.45	76.45
Nave Bayes + BOW	79.82	71.31
SVM + BOW	82.43	67.41
Random Forest + BOW	79.24	66.57
Linear Regression + BOW	77.45	64.90
Nave Bayes + HT	54.25	54.32
SVM + HT	49.75	47.63
Random Forest + HT	55.64	47.63
Linear Regression + HT	56.94	55.71

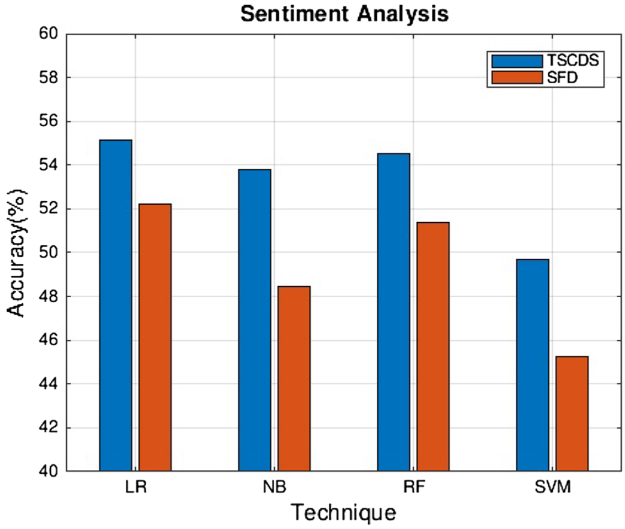


Fig. 3 Sentiment analysis with benchmark algorithm

respectively as shown in Table 3 and Fig. 3. In standalone class Linear Regression lead the performance approximate by 1.38%.

The performance of standalone classifier is significantly boast up when it is collaborated with POS for SA classification. Nave Bayes, SVM, Random Forest and Linear Regression gain 83.13, 83.27, 81.15, and 79.45% accuracy over SFD data set respectively as shown in Table 5 and Fig. 4a. In POS Class, SVM lead the performance approximate by 0.83% improvement. Whereas standalone classifier (Nave Bayes, SVM, Random Forest and Linear Regression) gain approximate 59.70, 73.38, 50.63 and 45.225 improvement over standalone class with POS as shown in Fig. 4b.

In BoW class, SVM serve better performance with 3.27% lead in twitter corpus. But with Stanford data set performance leaded by Navies Bayes with 5.78%. With BOW Nave Bayes, SVM, Random Forest and Linear Regression gain 79.82, 82.43, 79.24, and 77.45 accuracy over TSCDS and 71.31, 67.41, 66.57 and 64.90 accuracy over SFD data set respectively as shown in Table 5 and Fig. 5a. Whereas standalone classifier (Nave Bayes, SVM, Random Forest and Linear Regression) gain approximate 47.85, 57.82, 37.71 and 32.60 with BOW as shown in Fig. 5b.

In Hash tag class Linear Regression lead the performance approximate by 3.75% improvement. Nave Bayes, SVM, Random Forest and Linear Regression gain 54.25, 49.75, 55.64, and 56.94TSCDS and 54.32, 47.63, 53.34 and 55.71data set respectively as shown in Table 3 and Fig. 6a. Whereas standalone classifier (Nave Bayes, SVM, Random Forest and Linear Regression) gain approximate 6.21, 2.57, 2.92 and 4.93as shown in Fig. 6b.

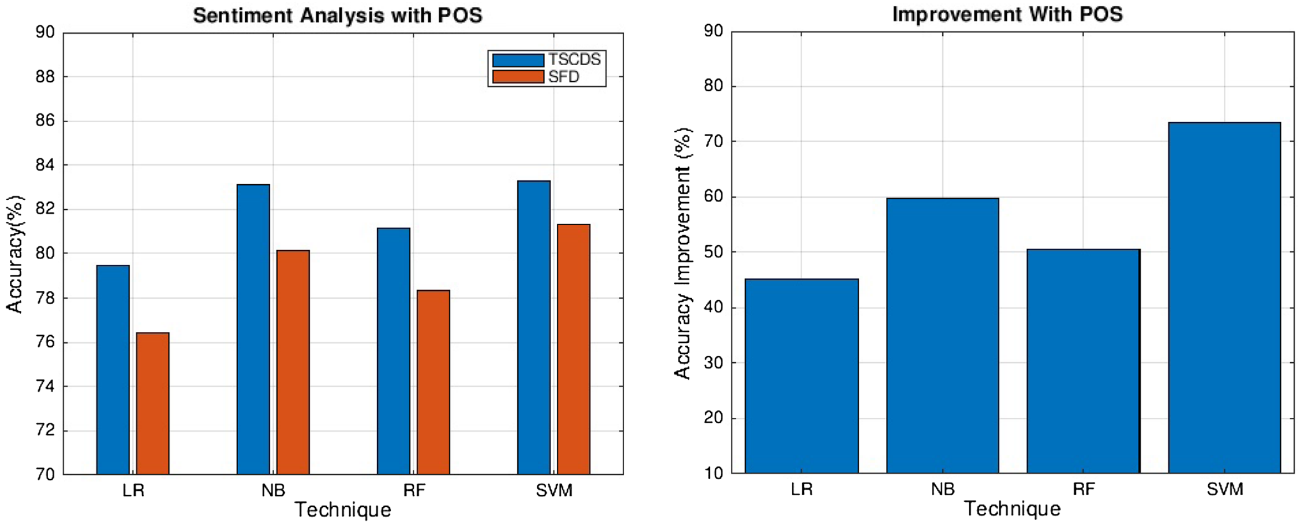
On other hand after analyzing the performance of classifier for SA with different feature extraction technique i.e. bag-of-words (BoW), feature hashing (HT), and Part of speech (POS) it is observed that classifiers gives better performance with POS feature extraction technique. Naive Bayes gain 83.19, 79.82 and 54.25 over TSCDS and 80.12, 71.31 and 54.32set with POS, BOW and HT respectively as shown in Fig. 7a. Navie Bayes classifier gain 59.70, 47.84 and 6.21standalone Nave Bayes shown in Fig. 7b.

Support vector machine gain 83.27, 82.43, 49.75over TSCDS and 81.34, 67.41 and 47.63set with POS, BOW and HT respectively as shown in Fig. 8a. SVM classifier gain 73.38, 57.82 and 2.57 improvement with POS, BOW and HT respectively w.r.t standalone SVM shown in Fig. 8b.

Random Forest gain 81.15, 79.24, 55.64and 78.34, 66.57 and 53.34POS, BOW and HT respectively as shown in Fig. 9a. RF classifier gain 50.63, 37.71 and 2.92BOW and HT respectively w.r.t standalone RF shown in Fig. 9b.

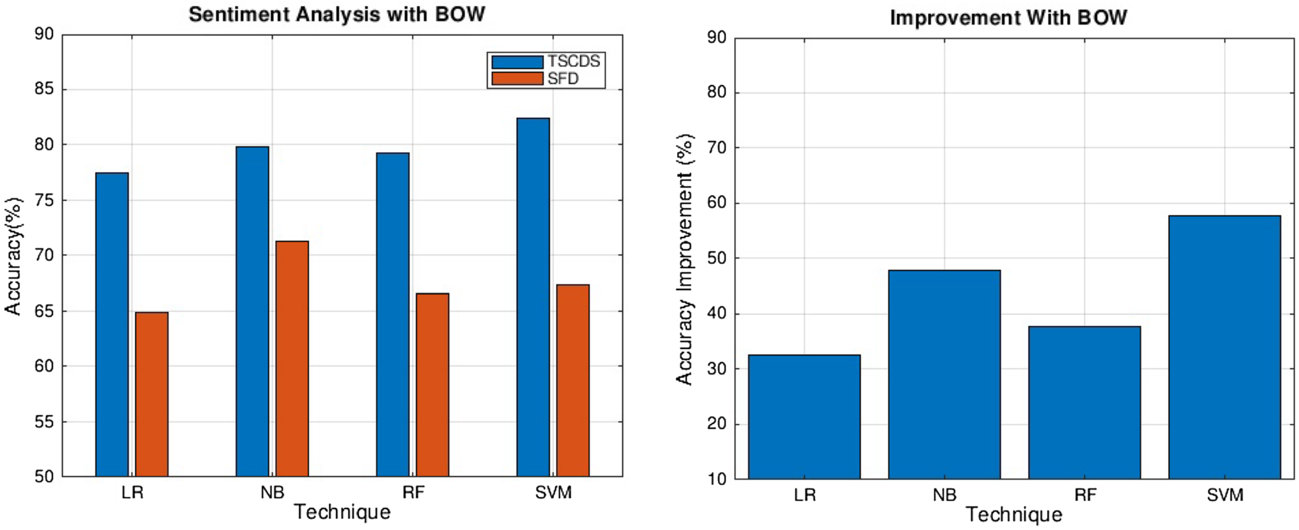
Linear Regression gain 79.45, 77.45 and 56.94TSCDS and 76.45, 64.90 and 55.71with POS, BOW and HT respectively as shown in Fig. 10a. LR classifier gain 45.22, 32.60 and 4.93POS, BOW and HT respectively w.r.t standalone LR shown in Fig. 10b.

The comparison of baseline sentiment classifier after identify the minimum and optimize set of feature vector following outcome has been acquired. POS is best suited feature extraction technique to identify the degree of dependency between feature value and labeled class over SA. Whereas BOW and Hass tagging gives biased result, SVM classifier comes out with best result with proposed framework for SA as compared to other classifier. Navie Bayes classifier inherently provides better result with POS



(a) Accuracy **(b)** Improvement Factor

Fig. 4 Cross comparison of SA technique using POS



(a) Accuracy **(b)** Improvement Factor

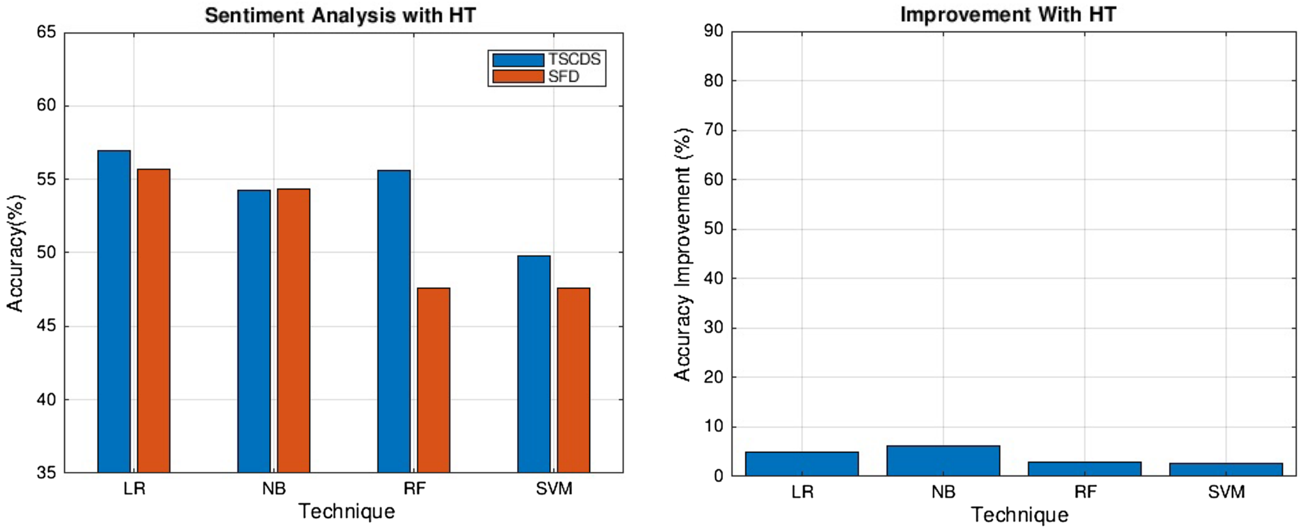
Fig. 5 Cross comparison of SA technique using BOW

and BOW feature extraction technique. Whereas Random Forest and linear regression provide the better result with Hass tagging.

7 Suggestion for future research

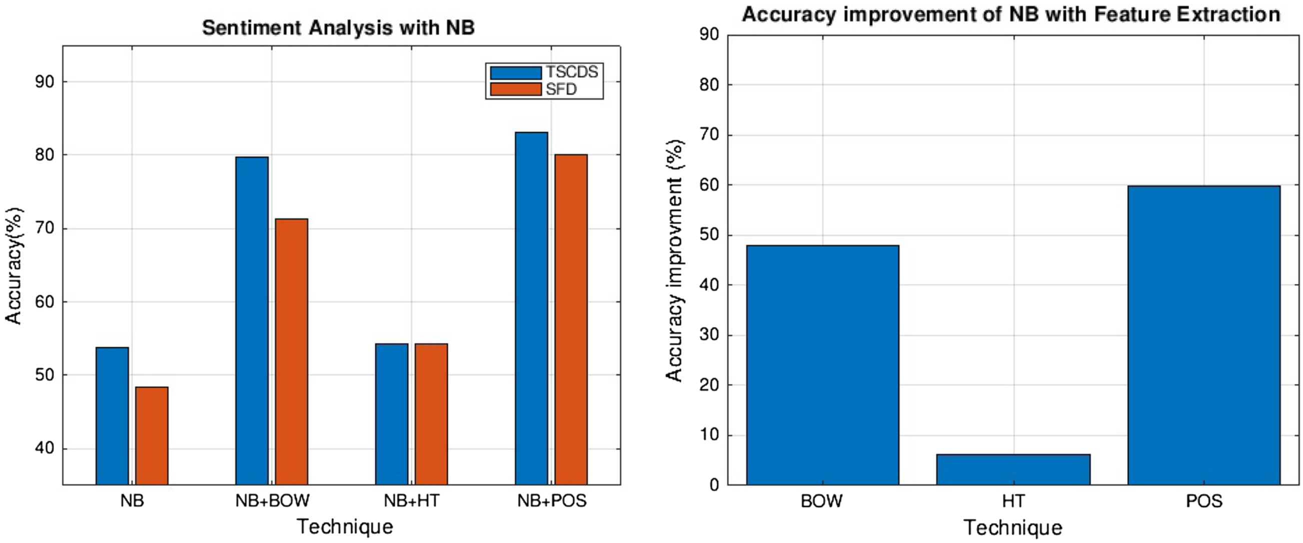
Recently huge amount of work has been carried out in field of sentiment analysis across the world namely China (Jin et al. 2007; Hogenboom et al. 2014), Germany (Smailovic et al. 2014), brazil (Abbasi and Chen 2007) and Arab Countries. But sentiment Analysis still has further research scope in order to create systems that can be reliably used in real-life applications.

(1) Sarcastic sentences: Sentiment analysis for sarcastic sentences still needs a lot of research. Consider a sentence S_6 contain negative complement with positive words. So S_6 should be classifying as positive but it is negative (Kranjc et al. 2015).



(a) Accuracy **(b)** Improvement Factor

Fig. 6 Cross comparison of SA technique using HT



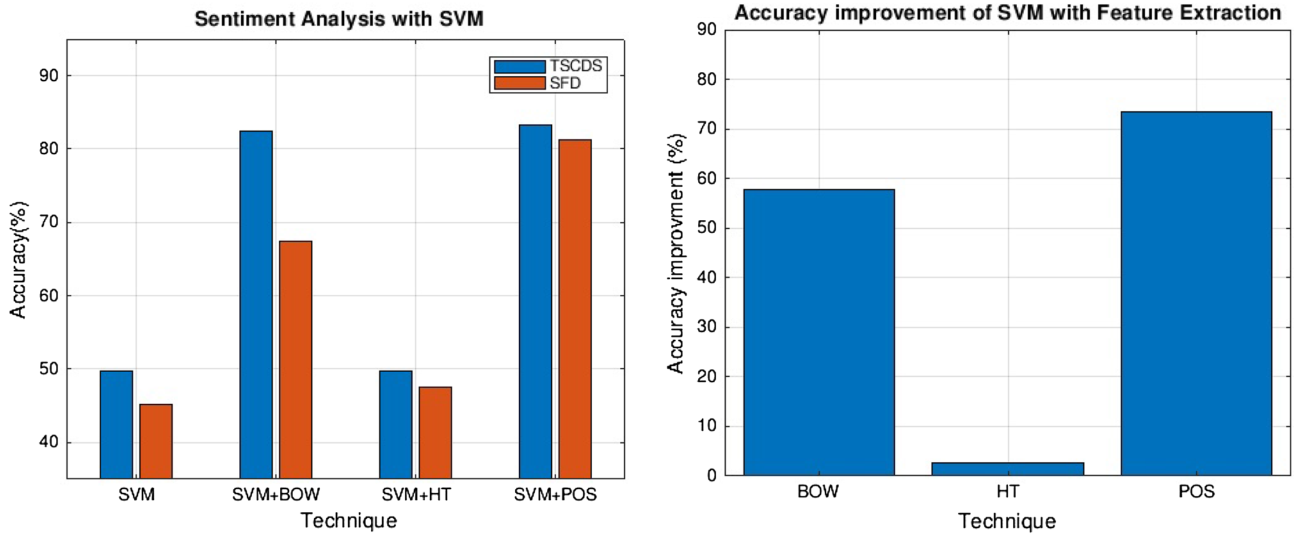
(a) Accuracy **(b)** Improvement Factor

Fig. 7 Cross comparison of feature extraction using NB over SA

S_6 : *Great idea, now try again with a real product development team.*

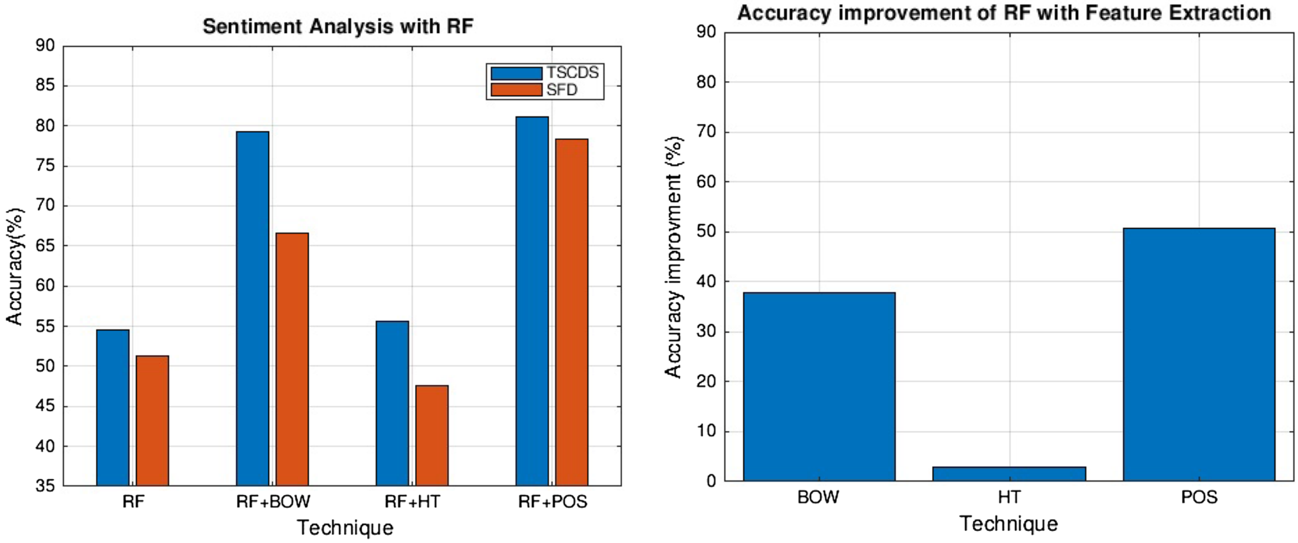
(2) Slangs, symbols, misspelled words and idioms: Recent Sentiment classifier fails to classify the sentence with slangs language, symbol, misspelled word and Idioms (Lau et al. 2014; Williams et al. 2015; Appel et al. 2016; Balahur and Perea-Ortega 2015; Balahur et al. 2014). Slang and Symbols terms are often only understood by like-minded person that can understand just what another person mean by using the latest slang terms. Consider a comment C_{13}, C_{14} and C_{15} contain slang, symbolic and idioms representative review over latest released I phone series.

C_{13} : *OMG! I cant believe how great this new I phone is.*



(a) Accuracy **(b)** Improvement Factor

Fig. 8 Cross comparison of feature extraction using SVM over SA



(a) Accuracy **(b)** Improvement Factor

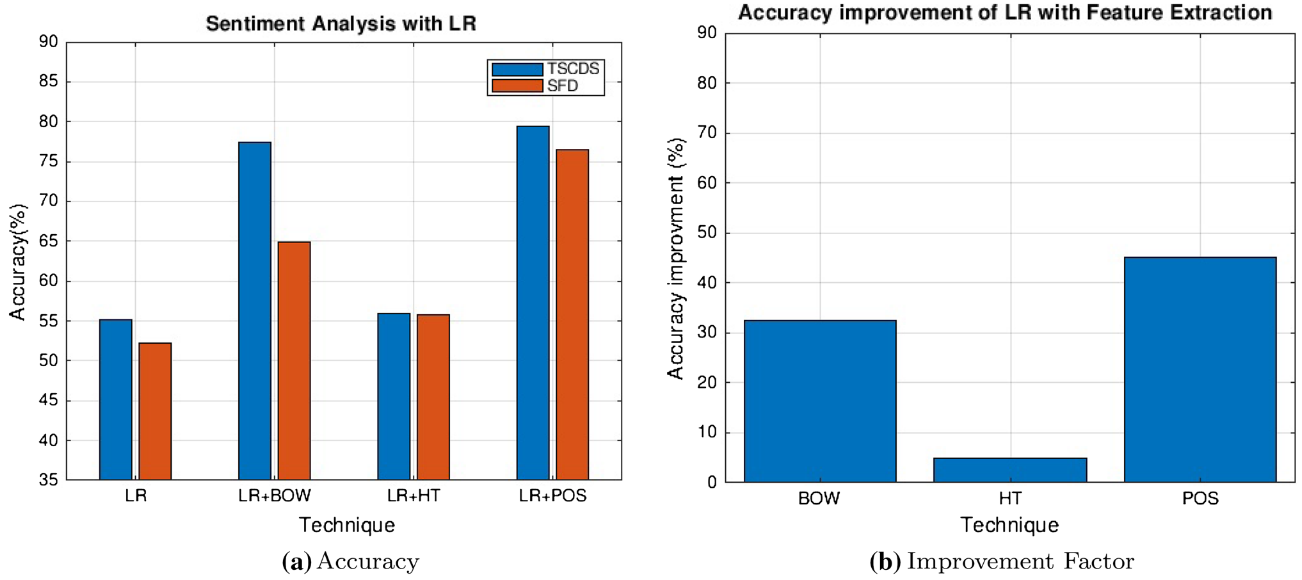
Fig. 9 Cross comparison of feature extraction using RF over SA

C_{14} :

😉👌😉👌😉👌

C_{15} : *I didnt exactly jump for joy. It is Costs an arm and a leg.*

Both C_{13} and C_{14} contain positive sentiment about I phone. Whereas person in Comment C_{15} is worried about I phone price and gives negative sentiment. For any automated system or unfamiliar person it is very tough to remind these notations, slang and Idioms. Whereas COCA (Corpus of Contemporary American English) (Davies 2018a), COHA (Corpus of Historical American English) (Davies 2018b), GloWbE (Davies 2018c) (Global Web-Based English Corpus) Corpus available for prepossessing the slang, idioms.



(a) Accuracy **(b)** Improvement Factor

Fig. 10 Cross comparison of feature extraction using LR over SA

(3) Annotated training data: Classification is an supervised learning approach. But for sentiment classification it has been noticed that there is lack of benchmark training data sets available.

(4) Sentiment strengths: Currently all the word tagged as positive or negative get the same score however, it would be possible to assign a different weight to different words, depending on the strength of the sentiment they transmit. For example consider the sentence S_7 and S_8 holds the positive polarity but S_7 would have a stronger influence than S_8.

(S_7) : *It is such a nice phone.*

(S_8) : *It is good phone.*

So it is need to assign a different weight to different words, depending on the strength of the sentiment they transmit (da Silva et al. 2014a).

(5) Multiple-language sentiment processing: Multiple languages post on social media platforms increase the complication of sentiment analysis with acceptable levels of accuracy and consistency. Recently some of research (Jin et al. 2007; Abbasi and Chen 2007; Cardie et al. 2006; Smailovic et al. 2014) classify the review for Chinese, Dutch, Spanish and Aurbi languages and few one for two level language (Kranjc et al. 2015; Hogenboom et al. 2014) like Chinese, English and German, English. Apart from this most of them are implement English as a target language.

(6) Significant Change towards negative sentiments: Some time sentiment of text depend upon position of negation occur. If a negation occurs near an adjective the polarity is estimated to the opposite of the polarity of adjective (Garca-Pablos et al. 2018; Ortigosa et al. 2014). For example consider the sentence S_9 should be classified as positive whereas S_{10} should be classified as negative.

Positive(S_9) : *Battery life is good.*

Negative(S_{10}) : *Battery life is not good.*

To classify this, the polarity is set to opposite of polarity of the occurring adjective when accompanied by "*not*". But if sentence S_{11} is consider, then sentence will be classify as negative whereas it need to be classify as positive.

Negative(S_{11}) : *I dont say battery life is not good.*

(7) Prediction time horizon: Recently researcher focus to analyze the trends of peoples sentiment along with time line and derived temporal sentiment analysis. The method extract topic trends with time stamps. For instance sentiment analysis for predicting stock exchange inflation need to be consider shrinking the granularity of the prediction time horizon which is to analyze the relationship between news impact and intra-day stock price return (Liu et al. 2003; Bravo-Marquez et al. 2014; Li et al. 2014).

8 Conclusion

Sentiment analysis is analyze people's sentiments, opinions, attitudes and emotions, towards an specific topics, products, individuals, organizations, and services. This paper present sentiment classification technique and summarized the recent research into supervised, unsupervised and hybrid SA technique. This work includes the comparison of different feature evaluation and classification techniques under SA. The comparison of features evaluation is carried out to identify the minimum and optimize set of feature vector. For the feature extraction POS is best suited as it identifies the degree of dependency between feature value and labeled class. On the other hand the BOW and Hass tagging gives biased result. In classification, SVM comes out with best result for SA as compared to Navie Bayes, Random Forest and Linear regression. Navie Bayes classification inherently provides better result but biasing with POS and BOW lag their performance behind SVM. Whereas in case of Hass tagging, Random Forest and Linear Regression provide the better result. Without any feature extraction prepossessing, Linear Regression has serve better performance than other over both the data set and leaded by approximate 1.38%. With POS feature extraction, SVM serve the better performance.Whereas in BoW class, SVM serve better performance with 2.37% lead in twitter corpus. But with Stanford data set performance leaded by Navies Bayes with 5.78%. This paper also address problems like the excessive simplicity while classifying, generally, only positive, negative or neutral categories are used. Along with that the incapability to aggregate ratings from different sentences or paragraphs, in order to get a general rate about a complete opinion. Handling idioms, symbols, misspelled words and sarcastic sentences is still a challenging task. Multiple languages post with geographical treatment on social media platforms increase the complication of sentiment analysis with acceptable levels of accuracy and consistency.

References

Abbasi A, Chen H (2007) Affect intensity analysis of dark web forums. In: 2007 IEEE intelligence and security informatics, pp 282–288. https://doi.org/10.1109/ISI.2007.379486

Appel O, Chiclana F, Carter J, Fujita H (2016) A hybrid approach to the sentiment analysis problem at the sentence level. Knowl Based Syst 108(Supplement C):110–124. https://doi.org/10.1016/j.knosys.2016.05.040 **(new avenues in knowledge bases for natural language processing)**

Bagheri A, Saraee M, de Jong F (2013) Care more about customers: unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowl Based Syst 52(Supplement C):201–213. https://doi.org/10.1016/j.knosys.2013.08.011

Balahur A, Perea-Ortega JM (2015) Sentiment analysis system adaptation for multilingual processing: the case of tweets. Inf Process Manag 51(4):547–556. https://doi.org/10.1016/j.ipm.2014.10.004

Balahur A, Turchi M (2014) Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput Speech Lang 28(1):56–75. https://doi.org/10.1016/j.csl.2013.03.004

Balahur A, Mihalcea R, Montoyo A (2014) Computational approaches to subjectivity and sentiment analysis: present and envisaged methods and applications. Comput Speech Lang 28(1):1–6. https://doi.org/10.1016/j.csl.2013.09.003

Bravo-Marquez F, Mendoza M, Poblete B (2014) Meta-level sentiment models for big social data analysis. Knowl Based Syst 69(Supplement C):86–99. https://doi.org/10.1016/j.knosys.2014.05.016

Cardie C, Farina C, Bruce T (2006) Using natural language processing to improve erulemaking: Project highlight. In: Proceedings of the 2006 international conference on digital government research. dg.o '06. Digital Government Society of North America, pp 177–178. https://doi.org/10.1145/1146598.1146651

Di Caro L, Grella M (2013) Sentiment analysis via dependency parsing. Comput Stand Interfaces 35(5):442–453. https://doi.org/10.1016/j.csi.2012.10.005

Carro RM, Ballesteros FJ, Ortigosa A, Guardiola G, Soriano E (2012) Angryemail? emotion-based e-mail tool adaptation. In: Proceedings of the 4th international conference on ambient assisted living and home care. Iwaal'12. Springer, Berlin, pp 399–406. https://doi.org/10.1007/978-3-642-35395-654.978-3-642-35394-9

Cho H, Kim S, Lee J, Lee JS (2014) Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. Knowled Based Syst 71(Supplement C):61–71. https://doi.org/10.1016/j.knosys.2014.06.001

Colace F, Casaburi L, De Santo M, Greco L (2015) Sentiment detection in social networks and in collaborative learning environments. Comput Hum Behav 51(Part B):1061–1067. https://doi.org/10.1016/j.chb.2014.11.090 **(computing for human learning, behaviour and collaboration in the social and mobile networks era)**

Cruz FL, Troyano JA, Pontes B, Javier Ortega F (2014) Building layered, multilingual sentiment lexicons at synset and lemma levels. Expert Syst Appl 41(13):5984–5994. https://doi.org/10.1016/j.eswa.2014.04.005

Cui Z, Shi X, Chen Y (2016) Sentiment analysis via integrating distributed representations of variable-length word sequence. Neurocomputing 187(Supplement C):126–132. https://doi.org/10.1016/j.neucom.2015.07.129 **(recent developments on deep big vision)**

da Silva NFF, Hruschka ER, Hruschka ER (2014) Tweet sentiment analysis with classifier ensembles. Decis Support Syst 66(C):170–179. https://doi.org/10.1016/j.dss.2014.07.003

da Silva NFF, Hruschka ER, Hruschka ER (2014) Tweet sentiment analysis with classifier ensembles. Decis Support Syst 66(Supplement C):170–179. https://doi.org/10.1016/j.dss.2014.07.003

da Silva NFF, Coletta LFS, Hruschka ER, Hruschka ER Jr (2016) Using unsupervised information to improve semi-supervised tweet sentiment classification. Inf Sci 355–356(Supplement C):348–365. https://doi.org/10.1016/j.ins.2016.02.002

Das S, Chen M (2001) Yahoo! for amazon: Extracting market sentiment from stock message boards. In: Asia pacific finance association annual conf. (APFA)

Davies M (2018a) Corpus of COCA. https://corpus.byu.edu/coca/. Accessed 1 Jan 2018

Davies M (2018b) Corpus of COHA. https://corpus.byu.edu/coha/. Accessed 1 Jan 2018

Davies M (2018c) Corpus of GloWbE. https://corpus.byu.edu/glowbe/. Accessed 1 Jan 2018

Deng Z-H, Luo K-H, Hong-Liang Y (2014) A study of supervised term weighting scheme for sentiment analysis. Expert Syst Appl 41(7):3506–3513. https://doi.org/10.1016/j.eswa.2013.10.056

Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56(4):82–89. https://doi.org/10.1145/2436256.2436274

Fernndez-Gavilanes M, Tl Lpez, Juncal-Martnez J, Costa-Montenegro E, Gonzlez-Castao FJ (2016) Unsupervised method for sentiment analysis in online texts. Expert Syst Appl 58(Supplement C):57–75. https://doi.org/10.1016/j.eswa.2016.03.031

Garca-Pablos A, Montse C, German R (2018) W2vlda: almost unsupervised system for aspect based sentiment analysis. Expert Syst Appl 91:127–137. https://doi.org/10.1016/j.eswa.2017.08.049

Ghiassi M, Skinner J, Zimbra D (2013) Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst Appl 40(16):6266–6282. https://doi.org/10.1016/j.eswa.2013.05.057

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision (CS224N–Final Project Report). Stanford University, Stanford, CA. http://www.yuefly.com/Public/Files/2017-03-07/58beb0822faef.pdf

Habernal I, Ptcek T, Steinberger J (2014) Supervised sentiment analysis in czech social media. Inf Process Manag 50(5):693–707. https://doi.org/10.1016/j.ipm.2014.05.001

Haddi E, Liu X, Shi Y (2013) The role of text pre-processing in sentiment analysis. Proc Comput Sci 17(Supplement C):26–32. https://doi.org/10.1016/j.procs.2013.05.005 (first international conference on information technology and quantitative management)

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90

Hogenboom A, Heerschop B, Frasincar F, Kaymak U, de Jong F (2014) Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decis Support Syst 62(Supplement C):43–53. https://doi.org/10.1016/j.dss.2014.03.004

Huang F, Zhang S, Zhang J, Yu G (2017) Multimodal learning for topic sentiment analysis in microblogging. Neurocomputing 253(Supplement C):144–153. https://doi.org/10.1016/j.neucom.2016.10.086 **(learning multimodal data)**

Jha V, Savitha R, Deepa Shenoy P, Venugopal KR, Sangaiah AK (2017) A novel sentiment aware dictionary for multi-domain sentiment classification. Comput Electr Eng 1:1. https://doi.org/10.1016/j.compeleceng.2017.10.015

Jin X, Li Y, Mah T, Tong J (2007) Sensitive webpage classification for content advertising. In: Proceedings of the 1st international workshop on data mining and audience intelligence for advertising, Adkdd '07, PP 28–33. New York, NY, USA. ACM. https://doi.org/10.1145/1348599.1348604.978-1-59593-833-6

Joshi D, Datta R, Fedorovskaya E, Luong QT, Wang JZ, Li J, Luo J (2011) Aesthetics and emotions in images. IEEE Signal Process Mag 28(5):94–115. https://doi.org/10.1109/MSP.2011.941851

Kang D, Park Y (2014) Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and vikor approach. Expert Syst Appl 41(4, Part 1):1041–1050. https://doi.org/10.1016/j.eswa.2013.07.101

Khan FH, Qamar U, Bashir S (2016) ESAP: a decision support framework for enhanced sentiment analysis and polarity classification. Inf Sci 367–368(Supplement C):862–873. https://doi.org/10.1016/j.ins.2016.07.028

Khan FH, Qamar U, Bashir S (2016b) Sentimi: Introducing pointwise mutual information with sentiwordnet to improve sentiment polarity detection. Appl Soft Comput 39:140–153. https://doi.org/10.1016/j.asoc.2015.11.016

Kontopoulos E, Berberidis C, Dergiades T, Bassiliades N (2013) Ontology-based sentiment analysis of twitter posts. Expert Syst Appl 40(10):4065–4074. https://doi.org/10.1016/j.eswa.2013.01.001

Kranjc J, Smailovic J, Podpecan V, Grcar M, Martin N, Nada L (2015) Active learning for sentiment analysis on data streams: methodology and workflow implementation in the clowdflows platform. Inf Process Manag 51(2):187–203. https://doi.org/10.1016/j.ipm.2014.04.001

Lau RYK, Li C, Liao SSY (2014) Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. Decis Support Syst 65(Supplement C):80–94. https://doi.org/10.1016/j.dss.2014.05.005 **(crowdsourcing and social networks analysis)**

Lei J, Rao Y, Li Q, Quan X, Wenyin L (2014) Towards building a social emotion detection system for online news. Future Gen Comput Syst 37(Supplement C):438–448. https://doi.org/10.1016/j.future.2013.09.024 **(special section: innovative methods and algorithms for advanced data-intensive computing special section: semantics, intelligent processing and services for big data special section: advances in data-intensive modelling and simulation special section: hybrid intelligence for growing internet and its applications)**

Li W, Guo K, Shi Y, Zhu L, Zheng Y (2018) Dwwp: domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. Knowl Based Syst 146:203–214. https://doi.org/10.1016/j.knosys.2018.02.004

Li X, Xie H, Chen L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowl Based Syst 69(Supplement C):14–23. https://doi.org/10.1016/j.knosys.2014.04.022

Liang J, Liu P, Tan J, Bai S (2014) Sentiment classification based on as-lda model. Proc Comput Sci 31(Supplement C):511–516. https://doi.org/10.1016/j.procs.2014.05.296 (2nd international conference on information technology and quantitative management, ITQM 2014)

Liu H, Lieberman H, Selker T (2003) A model of textual affect sensing using real-world knowledge. In: Proceedings of the 8th international conference on intelligent user interfaces. Iui '03, 125–132. New York, NY, USA. ACM. DOIurl10.1145/604045.604067

Liu SM, Chen J-H (2015) A multi-label classification based approach for sentiment classification. Expert Syst Appl 42(3):1083–1093. https://doi.org/10.1016/j.eswa.2014.08.036

Ma R, Wang K, Qiu T, Sangaiah AK, Lin Dan, Liaqat Hannan Bin (2017) Feature-based compositing memory networks for aspect-based sentiment classification in social internet of things. Future Gen Comput Syst 1:1. https://doi.org/10.1016/j.future.2017.11.036

Mahyoub Fawaz HH, Siddiqui Muazzam A, Dahab Mohamed Y (2014) Building an arabic sentiment lexicon using semi-supervised learning. J King Saud Univ Comput Inf Sci 26(4):417–424. https://doi.org/10.1016/j.jksuci.2014.06.003 (special issue on arabic NLP)

Martn-Valdivia M-T, Martnez-Cmara E, Perea-Ortega J-M, Alfonso Urea-Lpez L (2013) Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. Expert Syst Appl 40(10):3934–3942. https://doi.org/10.1016/j.eswa.2012.12.084

Mishne G, Glance NS (2006) Predicting movie sales from blogger sentiment. In: In AAAI spring symposium: computational approaches to analyzing weblogs, pp 155–158. http://dare.uva.nl/personal/search?identifier=4021a1e1-57d8-4943-a667-cae5ec339891

Montejo-Rez A, Daz-Galiano MC, Martnez-Santiago F, Urea-Lpez LA (2014a) Crowd explicit sentiment analysis. Knowl Based Syst 69(Supplement C):134–139. https://doi.org/10.1016/j.knosys.2014.05.007

Montejo-Rez A, Eugenio Martnez-Cmara M, Martn-Valdivia T, Alfonso Urea-Lpez L (2014b) Ranked wordnet graph for sentiment polarity classification in twitter. Comput Speech Lang 28(1):93–107. https://doi.org/10.1016/j.csl.2013.04.001

Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL (2015) Text mining of news-headlines for forex market prediction: a multi-layer dimension reduction algorithm with semantics and sentiment. Expert Syst Appl 42(1):306–324. https://doi.org/10.1016/j.eswa.2014.08.004

Ortigosa A, Martn JM, Carro RM (2014) Sentiment analysis in facebook and its application to e-learning. Comput Hum Behav 31(Supplement C):527–541. https://doi.org/10.1016/j.chb.2013.05.024

Poria S, Cambria E, Howard N, Huang G-B, Hussain A (2016) Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 174:50–59. https://doi.org/10.1016/j.neucom.2015.01.095

Ptaszynski M, Rzepka R, Araki K, Momouchi Y (2014) Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. Comput Speech Lang 28(1):38–55. https://doi.org/10.1016/j.csl.2013.04.010

Rill S, Reinel D, Scheidt J, Zicari RV (2014) Politwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowl Based Syst 69(Supplement C):24–33. https://doi.org/10.1016/j.knosys.2014.05.008

Rong W, Nie Y, Ouyang Y, Peng B, Xiong Z (2014) Auto-encoder based bagging architecture for sentiment analysis. J Vis Lang Comput 25(6):840–849. https://doi.org/10.1016/j.jvlc.2014.09.005 **(distributed multimedia systems DMS2014 part I)**

Smailovic J, Grcar M, Lavrac N, Nidaric M (2014) Stream-based active learning for sentiment analysis in the financial domain. Inf Sci 285(Supplement C):181–203. https://doi.org/10.1016/j.ins.2014.04.034 **(processing and mining complex data streams)**

Stepanov EA, Riccardi G (2011) Detecting general opinions from customer surveys. In: 2011 IEEE 11th international conference on data mining workshops, pp 115–122. https://doi.org/10.1109/ICDMW.2011.63

Vilares D, Gmez-Rodrguez C, Alonso MA (2017) Universal, unsupervised (rule-based), uncovered sentiment analysis. Knowl Based Syst 118:45–55. https://doi.org/10.1016/j.knosys.2016.11.014

Wang G, Zhang Z, Sun J, Yang S, Larson CA (2015) Pos-rs: a random subspace method for sentiment classification based on part-of-speech analysis. Inf Process Manag 51(4):458–479. https://doi.org/10.1016/j.ipm.2014.09.004

Williams L, Bannister C, Arribas-Ayllon M, Preece A, Spasic I (2015) The role of idioms in sentiment analysis. Expert Syst Appl 42(21):7375–7385. https://doi.org/10.1016/j.eswa.2015.05.039

Wu C-E, Tsai RT-H (2014) Using relation selection to improve value propagation in a conceptnet-based sentiment dictionary. Knowl Based Syst 69(Supplement C):100–107. https://doi.org/10.1016/j.knosys.2014.04.043

Xiao Z, Li X, Wang L, Yang Q, Jiayi D, Sangaiah AK (2017) Using convolution control block for chinese sentiment analysis. J Parallel Distrib Comput 1:1. https://doi.org/10.1016/j.jpdc.2017.10.018

Yan G, He W, Shen J, Tang C (2014) A bilingual approach for conducting Chinese and English social media sentiment analysis. Comput Netw 75(Part B):491–503. https://doi.org/10.1016/j.comnet.2014.08.021 (special issue on online social networks)

Yang J, She D, Sun M, Cheng MM, Rosin P, Wang L (2018) Visual sentiment prediction based on automatic discovery of affective regions. IEEE Trans Multimed PP(99):1–1. https://doi.org/10.1109/TMM.2018.2803520

Yu L-C, Wu J-L, Chang P-C, Chu H-S (2013) Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl Based Syst 41(Supplement C):89–97. https://doi.org/10.1016/j.knosys.2013.01.001

Zhang D, Hua X, Zengcai S, Yunfeng X (2015) Chinese comments sentiment classification based on word2vec and svmperf. Expert Syst Appl 42(4):1857–1863. https://doi.org/10.1016/j.eswa.2014.09.011

Zhang S, Wei Z, Wang Y, Liao T (2018) Sentiment analysis of chinese micro-blog text based on extended sentiment dictionary. Future Gen Comput Syst 81(Supplement C):395–403. https://doi.org/10.1016/j.future.2017.09.048

Zhou F, Jiao JR, Yang XJ, Lei B (2017) Augmenting feature model through customer preference mining by hybrid sentiment analysis. Expert Syst Appl 89(Supplement C):306–317. https://doi.org/10.1016/j.eswa.2017.07.021

Ziegelmayer D, Schrader R (2012) Sentiment polarity classification using statistical data compression models. In: 2012 IEEE 12th international conference on data mining workshops, pp 731–738. 1https://doi.org/10.1109/ICDMW.2012.43

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.