The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation

DAVID ZIMBRA, Santa Clara University
AHMED ABBASI, University of Virginia
DANIEL ZENG, University of Arizona
HSINCHUN CHEN, University of Arizona

Twitter has emerged as a major social media platform and generated great interest from sentiment analysis researchers. Despite this attention, state-of-the-art Twitter sentiment analysis approaches perform relatively poorly with reported classification accuracies often below 70%, adversely impacting applications of the derived sentiment information. In this research, we investigate the unique challenges presented by Twitter sentiment analysis, and review the literature to determine how the devised approaches have addressed these challenges. To assess the state-of-the-art in Twitter sentiment analysis, we conduct a benchmark evaluation of 28 top academic and commercial systems in tweet sentiment classification across five distinctive data sets. We perform an error analysis to uncover the causes of commonly occurring classification errors. To further the evaluation, we apply select systems in an event detection case study. Finally, we summarize the key trends and takeaways from the review and benchmark evaluation, and provide suggestions to guide the design of the next generation of approaches.

CCS Concepts: • **Information systems →Information systems applications; Data mining**; • **Computing methodologies →Artificial intelligence; Natural language processing;**

Additional Key Words and Phrases: Sentiment analysis, opinion mining, social media, twitter, benchmark evaluation, natural language processing, text mining

ACM Reference Format:
David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans. Mgmt. Info. Syst.*, 29 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Vast quantities of diverse user-generated social media are continuously produced including reviews, blogs, comments, discussions, images, and videos. These communications offer valuable opportunities to access and understand the perspectives of users on topics of interest, and contain information capable of explaining and predicting business and social phenomena like product sales [Liu 2006; Forman et al. 2008], stock returns [Das and Chen 2007; Zimbra et al. 2015], and the outcomes of political elections [Tumasjan et al. 2010; O'Connor et al. 2010]. Central to these analyses is the evaluation of the sentiment (opinion) expressed by

This work is supported by the National Science Foundation under grants IIS-1553109, IIS-1236970, BDS-1636933, CCF-1629450, and ACI-1443019, the MOST Grant 2016QY02D0305, the NNSFC Innovative Team Grant 71621002, the CAS Grant ZDRW-XH-2017-3, and the NIH Grant 5R01DA037378-04.

Author's addresses: D. Zimbra, Operations Management & Information Systems Department, Santa Clara University; A. Abbasi, Information Technology Area and Center for Business Analytics, University of Virginia; D. Zeng, Management Information Systems Department, University of Arizona; H. Chen, Artificial Intelligence Lab, University of Arizona.

DOI:http://dx.doi.org/10.1145/0000000.0000000

users in their text communications. Sentiment analysis is an active area of research motivated to improve the automated recognition of sentiment expressed in text, with performance gains leading to more effective application of the derived information.

Among the various social media platforms, Twitter has experienced particularly widespread user adoption and rapid growth in communication volume. Twitter is a micro-blogging platform where users author 'tweets' that are broadcasted to their followers or sent to another user. As of 2016, Twitter has over 313 million users active within a given month, including 100 million users daily [Twitter 2016]. They are distributed globally, with 77% located outside of the US, generating over 500 million tweets per day [Twitter 2014]. The Twitter website ranks 8th globally for traffic [Alexa 2015], and responds to over 15 billion API calls daily [DuVander 2012]. Twitter content also appears in over 1 million third-party websites [Twitter 2014].

Accompanying this tremendous growth, Twitter has also been the subject of much recent sentiment analysis research, as tweets often express a user's opinion on a topic of interest. Tweets have provided valuable insights on issues related to business and society [Jansen et al. 2009; Gleason 2013]. Researchers have utilized information derived through Twitter sentiment analysis (TSA) to explain and predict product sales [Rui et al. 2013], stock market movements [Bollen et al. 2011], and the outcomes of political elections [Bermingham and Smealton 2010]. The Twitter platform enables the quick distribution of information across a large population of users, and is highly effective for disseminating news in real time. To capture this emerging information, researchers have developed approaches to monitor Twitter and detect various events, such as shifts in the public opinion regarding presidential candidates [Wang et al. 2010], indications of financial market movements [Zhang and Skiena 2010], or early warnings of adverse medical events [Abbasi et al. 2013].

These and other applications of information derived through TSA are dependent upon their underlying approaches to evaluating the sentiments expressed by users in their tweets. Consequently, much research has been devoted to developing improved approaches to TSA. Despite this attention, and a growing body of literature, the performances of state-of-the-art TSA approaches remain poor with reported tweet sentiment classification accuracies below 70% [Hassan et al. 2013]. These lackluster performances may be attributed to several characteristics of tweets that make TSA challenging, particularly when compared with other genres of communication. Considering the popularity of Twitter, and recent studies that have demonstrated the value of information derived through TSA while revealing the difficulties experienced by TSA approaches, a thorough investigation is warranted.

Therefore, we examine the following research questions:

- What are the areas and motivations for TSA research?
- What are the challenges presented by TSA?
- How have TSA approaches addressed these challenges?
- How will state-of-the-art TSA approaches perform in a benchmark evaluation in tweet sentiment classification?
    - What are the causes of commonly occurring classification errors?
    - How will select approaches perform when applied in Twitter monitoring and event-detection?
- What are the key trends and takeaways from the TSA review and evaluation?

To address these research questions, we first briefly introduce TSA, describe commonly applied approaches and major motivations for recent research. We then review the TSA literature, discuss the unique challenges associated with TSA, and present a taxonomy of the techniques devised in prior studies to address these

challenges. To assess the state-of-the-art in TSA, we then conduct a benchmark evaluation of 28 top academic and commercial systems in tweet sentiment classification across five distinctive Twitter data sets. Following the experiments, we perform an error analysis to uncover the root causes of commonly occurring classification errors made by the systems. Since TSA systems are often deployed to monitor Twitter and detect the occurrences of specific events, we then apply the top-performing systems in an event detection case study. Finally, we summarize the key trends and takeaways from the review and benchmark evaluation, and provide suggestions to guide the design of the next generation of TSA approaches.

2. TWITTER SENTIMENT ANALYSIS

Twitter sentiment analysis is a specialized area within sentiment analysis, a prominent topic of research in the field of computational linguistics. Approaches to sentiment analysis identify and evaluate opinions expressed in text using automated methods. Sentiment analysis has been performed in a variety of genres of communication including professional media like news articles [Tetlock 2007] as well as social media like product reviews [Pang et al. 2002; Dave et al. 2003; Gamon 2004], web forums [Das and Chen 2007; Abbasi et al. 2008], and Facebook [Troussas et al. 2013; Ortigosa et al. 2014]. The growth in sentiment analysis research has followed that of social media, as researchers and firms pursue the valuable opinions of large populations of users.

Sentiment analysis tasks include classification of sentiment polarity expressed in text (e.g., positive, negative, neutral), identifying sentiment target/topic, opinion holder identification, and identifying sentiment for various aspects of a topic, product, or organization [Abbasi et al. 2008]. Sentiment polarity classification has emerged as one of the most studied tasks due to its significant implications for various social media analytics use cases. The sentiment polarity classification problem is often modeled as a two-way (positive/negative) or three-way (positive/negative/neutral) classification of a unit of text, although some methods produce more fine-grained classifications or continuous intensity scores. For example, researchers have evaluated five-sentiment-class models in brand-related TSA to target strong and mild positive and negative sentiments that provide more actionable intelligence to brand management practitioners [Jansen et al. 2009; Ghiassi et al. 2013; 2016]. Sentiment polarity classifications may also be specific to the domain of analysis. For example, tweets classified as positive may express preference for a candidate in a political TSA [Mejova 2013], optimism about a firm's future financial performance in a stock TSA [Smailovic 2013], satisfaction with a product in a marketing TSA [Rui et al. 2013], or the effectiveness of a drug in treating an ailment in a medical TSA [Androver et al. 2014]. Sentiment analysis is performed at various units of text, from phrases and sentences, to messages and entire documents.

Approaches devised for TSA typically follow those developed for more traditional genres of communication and other social media like product reviews and web forums, and can be broadly categorized into two classes. The first class involves the use of a lexicon of opinion-related terms with a scoring method to evaluate sentiment in an unsupervised application [Turney 2002; Kim and Hovy 2004]. These methods are widely applicable but their performances are limited as they are unable to account for contextual information, novel vocabulary, or nuanced indicators of sentiment expression. The second class quantify the text based upon a feature representation, and apply a machine learning algorithm to derive the relationship between feature values and sentiment using supervised learning [Pang et al. 2002; Gamon 2004].

Models based upon supervised learning require a large set of training instances with sentiment class labels to calibrate model parameters, and training domain specificity limits their potential for broader application.

To collect the TSA literature for our review, we utilized three approaches: citation analysis, keyword search, and browsing. We first focused on developing a seed set of early TSA publications to use in our citation analysis. Using the Google Scholar academic search engine, we searched for journal or conference publications containing any Twitter-related keywords in their titles (e.g., Twitter, tweet, microblog), published in the first few years following the founding of Twitter in 2006. We carefully scrutinized each of the retrieved publications that satisfied our search criteria for inclusion in our seed set of TSA literature. To supplement the search engine, we also browsed the archives of Information Systems and Computer Science journals and conferences where sentiment analysis and social media-related studies were likely to appear in these years, including for example the journals ACM Transactions on Information Systems, IEEE Transactions on Knowledge and Data Engineering, and the Journal of the American Society for Information Science and Technology, and the proceedings of the AAAI Conference on Web and Social Media, Association for Computational Linguistics Conference, ACM Conference on Information and Knowledge Management, ACM Conference on Web Search and Data Mining, International Conference on Computational Linguistics, Conference on Empirical Methods in Natural Language Processing, and the International World Wide Web Conference, among others. Articles focused on TSA began appearing in the literature in 2009 and 2010. We formed the seed set of studies for citation analysis using TSA publications appearing in these two years: Go et al. [2009], Jansen et al. [2009], Barbosa and Feng [2010], Bermingham and Smeaton [2010], Bifet and Frank [2010], Davidov et al. [2010], Diakopoulos and Shamma [2010], O'Connor et al. [2010], Pak and Paroubek [2010a; 2010b], Thelwall et al. [2010], Tumasjan et al. [2010], and Zhang and Skiena [2010]. Our citation analysis had two directions of search: evaluating earlier studies cited by the publications in our collection, and evaluating studies published more recently that cited the publications in our collection. First, we carefully scrutinized the list of citations of each publication in our seed set for additional studies to include. Then, we utilized Google Scholar to retrieve the more recent publications that cited each of the publications in our seed set. Studies identified through these approaches with titles containing Twitter or sentiment analysis keywords, or published in journals or conferences where related research appeared were acquired and evaluated for inclusion. If additional relevant studies were identified and added to our collection through the citation analysis, we reiterated the citation analysis procedure focusing on these newly introduced publications. Citation analysis continued until no additional relevant TSA literature was identified to add to our collection.

Following citation analysis, we expanded our keyword search and browsing approaches to identify any additional relevant TSA literature. We also utilized general keywords in our Google Scholar searches related to social media (e.g., social media, message, blog, post, etc.) and sentiment analysis (e.g., sentiment, opinion, sentiment analysis, opinion mining, etc.) in addition to Twitter-related keywords, and searched for studies published any time after the founding of Twitter in 2006. We also carefully browsed the archives of journals and conferences where TSA studies were likely to appear after 2006. In addition to the journals and conferences previously listed, we utilized the publications identified for our collection through citation analysis and keyword search to inform our browsing. When an added

publication introduced a new journal or conference to our collection, this journal or conference was then targeted for our archival browsing. In total, the archives of more than 10 journals and 20 conferences were browsed for TSA literature to add to our collection. While we do not consider our collection of TSA literature to be exhaustive, through our systematic approaches to citation analysis, keyword search, and browsing, we have identified a broad and representative collection of TSA literature for our review consisting of more than 70 TSA publications.

TSA research follows two major motivations. The first line of research focuses on the *application* of TSA to gain insights into various business or social issues, predict key indicators, or monitor Twitter for emerging information or events. Recognizing the value of information derived through accurate TSA, the second line of research focuses on innovating and developing improved *techniques* for TSA. These two lines of research are inherently related, and motivate the advancement of one another, as performance gains in TSA techniques lead to more effective application of the derived Twitter sentiment information. Improvements in TSA techniques translate to clearer insights regarding issues of interest, greater accuracy in predicting key indicators related to social sentiments, and faster detection of emerging events.

These motivations have similarly advanced sentiment analysis research in other social media. The sentiment analysis techniques developed in studies on product reviews [Pang et al. 2002], web forums [Abbasi et al. 2008], or Facebook comments [Troussas et al. 2013] are often intended to improve the quality of information derived from these sources when applied to gain insights, into for example product sales [Forman et al. 2008] or stock prices [Das and Chen 2007; Siganos et al. 2014].

We next describe research on the application of TSA, including descriptive case studies and studies focused on event detection or prediction using Twitter. We then review research on technical studies aimed at developing improved methods for TSA. We describe the unique challenges associated with TSA, and develop a taxonomy for the techniques devised in prior studies specifically to address these challenges.

2.1 Twitter Sentiment Analysis Applications

TSA has been applied effectively in descriptive case studies to improve the understanding of user opinion on diverse business and social issues, like a product brand [Jansen et al. 2009], presidential candidate performances in a debate [Diakopoulos and Shamma 2010] or primary election [Mejova 2013], supreme court decision [Clark et al. 2014], nuclear power generation [Kim and Kim 2014], the holiday season [Hu 2013], and patient reactions to medicines [Adrover et al. 2014]

Sentiments derived through TSA have also been useful in explaining and predicting key business and social indicators, such as stock market movements [Bollen et al. 2011; Mittal and Goel 2012; Smailovic et al. 2013], product sales [Rui et al. 2013; Verma et al. 2015], and the outcomes of political elections [Tumasjan et al. 2010; O'Connor et al. 2010; Bermingham and Smealton 2010; Chung and Mustafaraj 2011; Gayo-Avello 2013; Ringsquandl and Petkovic 2013].

Researchers have performed TSA to monitor Twitter for fluctuations associated with specific events. Thelwall et al. [2011] identified significant changes in public opinion surrounding several business and social events. Wang et al. [2010] monitored shifts in sentiment regarding candidates in the 2012 presidential election. Researchers have devised stock-trading methods based upon TSA [Zhang and Skiena 2010; Rao and Srivastava 2014]. Public health and epidemic outbreaks [Ji et al. 2013] and other adverse medical events [Hassan et al. 2013; Abbasi and Adjeroh 2014; Adjeroh et al. 2014; Sharif et al. 2014] have also been monitored.

2.2 Twitter Sentiment Analysis Techniques

The effectiveness of these and other applications of information derived through TSA are critically dependent upon their underlying approach to evaluating the sentiments expressed by users in their tweets. Many TSA approaches are directly adopted from the literature on more established social media like product reviews and web forum messages. However, several characteristics of tweets complicate the analysis and challenge TSA approaches, resulting in generally poor sentiment classification performance with accuracies often below 70% [Hassan et al. 2013]. These challenging characteristics include the brevity of tweets and resulting compact, novel language with Twitter-specific communication elements [Bermingham and Smeaton 2010; Ghiassi et al. 2013], a strong sentiment class imbalance [Hagen et al. 2015], and stream-based tweet generation [Vanzo et al. 2014; Amati et al. 2014]. Other reviews of the TSA literature have also cited these challenges [Martinez-Camara et al. 2012; Bhuta et al. 2014; Giachanou and Crestani 2016]. Attention to these three challenges is required to achieve accurate TSA, and generate the benefits associated with effective application of the derived information.

Similar challenges have also affected sentiment analysis in other social media to varying degrees. For example, casual communications with frequent use of slang and acronyms are prevalent in social media [Pang et al. 2002; Dave et al. 2003; Gamon 2004; Das and Chen 2007; Abbasi et al. 2008], but the extreme length restriction applied to tweets intensifies this communication behavior and promotes the development of a compact, novel language among Twitter users that is particularly difficult to analyze [Hassan et al. 2013]. For these reasons some sentiment analysis researchers have avoided performing TSA and instead opted to focus on Facebook status updates which have up to 5000 characters [Troussas et al 2013]. Other social media are also characterized by a strong sentiment class imbalance, but the nature of the imbalance varies among social media. For example, product reviews are predominantly positive or negative [Pang et al. 2002; Dave et al. 2003], while in web forums neutral messages are more frequently observed [Abbasi et al. 2008]. While we focus solely on TSA in this review, the approaches devised in the literature to address the challenging characteristics of tweets and improve TSA performance may also be relevant or applicable in other social media.

Through our review of the literature, we have developed a taxonomy of the techniques devised to address these three unique challenges associated with TSA. Prior TSA reviews have focused primarily on the features utilized to represent tweets, classifiers and analysis methods, evaluation metrics, and TSA data sets [Martinez-Camara et al. 2012; Bhuta et al. 2014; Giachanou and Crestani 2016]. In our review, we focus directly on the unique challenges associated with TSA and the approaches and techniques devised by researchers to address them specifically. Eight distinctive classes of techniques were identified. The techniques of sentiment information propagation, feature representation expansion, Twitter-specific preprocessing, and Twitter-specific features attempt to address the TSA challenge of tweet brevity and novel Twitter language. Training set expansion, multiple classifier methods, and sentiment-topic model techniques address the TSA challenge of sentiment class imbalance and poor sentiment recall. Stream-based classifiers address the TSA challenge of stream-based tweet generation and temporal dependency. In Table I we list the challenging characteristics associated with TSA, the techniques designed specifically to address these challenges, a brief description of each technique, and representative studies.

Table I. Taxonomy of Techniques to Address Twitter Sentiment Analysis Challenges

TSA Challenge	TSA Technique	Technique Description	Representative Studies
Tweet Brevity; Novel Twitter Language	Sentiment Information Propagation	Propagation of known sentiment information throughout tweets to identify novel expressions of sentiment	[Cui et al. 2011; Mittal & Goel 2012; Tang et al. 2014; Zhou et al. 2014; Dong et al. 2014; Saif et al. 2014a; Kaewpitakkun et al. 2014; Hu et al. 2013]
	Feature Representation Expansion	Expansion of the tweet feature representation by supplementing or forming additional combinations of the tweet contents	[Saif et al. 2014a; Montejo-Raez et al. 2014; Sharif et al. 2014; Konotopoulos et al. 2013; Jiang et al. 2014]
	Twitter-Specific Preprocessing	Removal, replacement, or correction of Twitter-specific features like emoticons, hashtags, hyperlinks, user mentions, acronyms, or slang	[Bermingham & Smealton 2010; Kontopoulos et al. 2013; Dong et al. 2014; Kaewpitakkun et al. 2014; Montejo-Raez et al. 2014; Pak & Paroubek 2010a; 2010b; Zhang et al. 2011; Tan et al. 2012; Agarwal et al. 2011; Kouloumpis et al. 2011; Liu et al. 2012; Vanzo et al. 2014; Hu et al. 2013; Xiang and Zhou 2014; Saif et al. 2012; Saif et al. 2014b; Bakliwal et al. 2013; Ghiassi et al. 2013; Jiang et al. 2014; Aston et al. 2014]
	Twitter-Specific Features	Incorporation of Twitter-specific features like emoticons, hashtags, hyperlinks, acronyms, or slang into the tweet feature representation	[Cui et al. 2011; Hu et al. 2013; Zhang et al. 2013; Khan et al. 2014; Davidov et al. 2010; Agarwal et al. 2011; Kouloumpis et al. 2011; Liu et al. 2012; Vanzo et al. 2014; Hu et al. 2013; Xiang & Zhou 2014; Barbosa & Feng 2010; Bakliwal et al. 2013; Nielsen 2011; Ghiassi et al. 2013]
Sentiment Class Imbalance; Poor Sentiment Recall	Training Set Expansion	Expansion of the number of tweets available for training by considering noisy sentiment class labels	[Tang et al. 2014; Cui et al. 2011; Montejo-Raez et al. 2014; Jiang et al. 2011; Pak & Paroubek 2010a; 2010b; Zhang et al. 2013; Go et al. 2009; Bifet and Frank 2010; Kouloumpis et al. 2011; Liu et al. 2012; Xiang & Zhou 2014; Saif et al. 2012; Barbosa & Feng 2010; Speriosu et al. 2011; Jiang et al. 2014]
	Multiple Classifier Methods	Multiple tweet sentiment classifiers composed in an ensemble or multi-stage classification scheme	[Mittal & Goel 2012; Kontopoulos et al. 2013; Jiang et al. 2011; Khan et al. 2014; Davidov et al. 2010; Liu et al. 2012; Tapia & Velasquez 2014; Barbosa & Feng 2010; Aston et al. 2014]
	Sentiment-Topic Models	Evaluation of the topics discussed and the sentiments expressed regarding these identified topics in an integrated model	[Dong et al. 2014; Tan et al. 2012; Hu et al. 2013; Xiang & Zhou 2014; Saif et al. 2012; Huang et al. 2013; Amati et al. 2014]
Stream-Based Generation; Temporal Dependency	Stream-Based Classifiers	Sentiment classification of a stream of tweets through sequence labeling or proportional sentiment estimation	[Bifet & Frank 2010; Vanzo et al. 2014; Hu et al. 2013; Aston et al. 2014; Amati et al. 2014]

2.2.1. Techniques to Address Tweet Brevity and Novel Twitter Language. Tweets are brief messages, limited to 140 characters in length. This brevity greatly impacts the performance of TSA approaches, as there are relatively few terms to evaluate and score using a sentiment lexicon, or tweet feature representation vectors are very

sparsely populated [Hassan et al. 2013]. The length limitation also motivates users to create novel forms of expression, developing compact, casual language with frequent use of slang, acronyms, and emoticons. These novel terms may be unknown to the TSA approach at analysis, as the Twitter language is diverse and quickly evolving.

One class of TSA techniques addresses this challenge by propagating known sentiment information throughout tweets to identify new expressions of sentiment. Beginning with a seed set of emoticons or a lexicon of sentiment terms, these approaches propagate sentiment information from these known terms to other words used in the tweets, based upon their co-occurrence or proximity [Tang et al. 2014], [Zhou et al. 2014; Dong et al. 2014; Saif et al. 2014; Kaewpitakkun et al. 2014]. Once propagated, a larger vocabulary of terms with assigned sentiment value is generated, which may improve recall and precision in TSA. In related techniques, Cui et al. [2011] used emoticons to perform multilingual sentiment analysis, Kaewpitakkun et al. [2014] propagated sentiment intensity into the feature weighting of an SVM, and Hu et al. [2013] utilized a graph model to uncover the sentiment of unknown terms.

Another class addresses tweet brevity by expanding the tweet feature representation, supplementing the tweet or forming additional combinations of its contents. Some approaches reference lexical resources like WordNet to add word synonyms, hypernyms, and antonyms [Montejo-Raez et al. 2014; Sharif et al. 2014]. Others develop domain ontologies [Kontopoulos et al. 2013], or add semantic concept terms related to named entities [Saif et al. 2014]. Other techniques utilize novel combinations of the tweet contents to include in the feature representation, like skip grams [Fernandez et al. 2014], or target-dependent features [Jiang et al 2011].

Twitter has unique communication elements that offer novel ways of expressing sentiment, increasing the diversity of the Twitter language, and further complicating the TSA. User mentions (e.g. @username) direct the tweet to the mentioned user but also serve as a reference to the user. Hashtags (e.g. #hashtag) are included to link a tweet to others on the same topic, and allow users to find it easily. While hashtags were originally intended as topical markers, they often express more complex concepts and can convey sentiment information. Hyperlinks may also be included in a tweet, linking to websites containing information a user wishes to share. These Twitter-specific communication elements are significant to users and have specific meaning, which may not be represented entirely in the content of the tweet. Broader communal knowledge is often referenced through user mentions and hashtags.

One class of TSA techniques addresses the challenges associated with these Twitter-specific communication elements by removing, replacing, or correcting them in a preprocessing procedure prior to sentiment analysis. This preprocessing intends to reduce the diversity of the Twitter language and alleviate feature sparsity, and also allow for the application of lexical resources developed for more traditional genres of communication. Some approaches removed any occurrences of hashtags, hyperlinks, or user mentions [Kaewpitakkun et al. 2014; Montejo-Raez et al. 2014; Kontopoulos et al. 2013; Pak and Paroubek 2010a; 2010b; Zhang et al. 2011; Tan et al. 2012]. Others replaced them with common tokens [Bermingham and Smealton 2010; Dong et al. 2014; Liu et al. 2012; Tapia and Velasquez 2014; Vanzo et al. 2014]. Preprocessing procedures have also been devised for slang, acronyms, emoticons, and stylistic exaggerations. Some researchers have corrected stylistic exaggerations and misspellings [Mittal and Goel 2012; Kaewpitakkun et al. 2014; Jiang et al. 2011; Go et al. 2009; Bifet and Frank 2010; Xiang and Zhou 2014]. Others have replaced slang, emoticons, and acronyms with their proper word equivalents [Mittal and Goel 2012; Zhang et al. 2011; Tan et al. 2012; Xiang and Zhou 2014]. Infrequently occurring

topics or named entities have also been replaced with their associated semantic concepts [Bermingham and Smealton 2010; Sharif et al. 2014; Saif et al. 2012]. Stop word preprocessing procedures have been examined in detail [Saif et al. 2014b].

Alternatively, another class incorporates these Twitter-specific elements as well as slang, acronyms, and emoticons into the tweet feature representation. Hashtags [Zhang et al. 2011; Davidov et al. 2010; Agarwal et al. 2011; Bakliwal et al. 2013] and hyperlinks [Agarwal et al. 2011; Barbosa and Feng 2010; Bakliwal et al. 2013] are commonly included. Acronyms and slang are prevalent and often convey sentiments, and have also been considered [Agarwal et al. 2011; Kouloumpis et al. 2011; Hu et al. 2013; Barbosa and Feng 2010; Bakliwal et al. 2013; Ghiassi et al. 2013]. Emoticons represent rich sentiment information and are often featured [Mittal and Goel 2012; Cui et al. 2011; Jiang et al. 2011; Zhang et al. 2011; Ghiassi et al. 2013].

2.2.2. Techniques to Address Sentiment Class Imbalance and Poor Sentiment Recall. Tweets are predominantly neutral with relatively few expressing either positive or negative sentiment, resulting in a highly imbalanced multi-class classification problem [Hagen et al. 2015]. This is in contrast to other social media where sentiment analysis is performed, such as product reviews, where positive or negative opinions are regularly expressed while neutral comments are infrequent. A large sentiment class imbalance is problematic as it may bias machine-learned models, inducing a preference for the neutral class, and resulting in low recall of the sentiment classes.

One class of TSA techniques attempts to address the challenges associated with the sentiment class imbalance by expanding the training set utilized to calibrate the machine-learned classifiers. A larger training set provides greater exposure to the various expressions of sentiment used in tweets, and may improve the recall of positive and negative sentiment classes. A widely used approach is to consider emoticons as noisy class labels [Tang et al. 2014; Montejo-Raez et al. 2014; Jiang et al. 2011; Pak and Paroubek 2010; Go et al. 2009; Jiang et al. 2014]. Others have similarly considered hashtags [Davidov et al. 2010; Kouloumpis et al. 2011]. The training set has also been expanded to include topically-related tweets [Jiang et al. 2011] or tweets of followers [Speriosu et al. 2011]. In related techniques, Xiang and Zhou [2014] iteratively expanded the training set using model classifications, and Liu et al. [2012] used an emoticon-labeled data set to smooth classifier parameters.

To improve sentiment recall and address the class imbalance, another class of TSA techniques applies multiple sentiment classifiers in an ensemble or multi-stage classification scheme. Since sentiment classifier performance varies considerably, multi-classifier ensembles have been developed to integrate the strengths of the individual constituent approaches [Bravo-Marquez et al. 2013; Goncalves et al. 2013]. Multi-stage classification schemes have also been devised, applying a subjectivity classifier first, then a sentiment classifier [Jiang et al. 2011; Tapia and Velasquez 2014; Aston et al. 2014]. Mittal and Goel [2012] and Khan et al. [2014] developed three-stage classification schemes where early stages captured obvious indicators of sentiment while more nuanced expressions were evaluated in later stages. Other researchers have examined phrase or entity-level classification [Speriosu et al. 2011].

A related class improves the recall of sentiments regarding a specific topic using integrated sentiment-topic models. Most applications of TSA aim to measure the opinion of users regarding a specific product, company, person, or issue. Sentiment-topic models are designed to capture the sentiment pertaining to a targeted topic, isolated from other sentiments that may be expressed in the tweet. For example, researchers have devised integrated mixed models based upon latent dirichlet

allocation (LDA) [Xiang and Zhou 2014; Saif et al. 2012; Huang et al. 2013]. In a related technique, Tan et al. [2012] first performed topical analysis using LDA before sentiment analysis. Dong et al. [2014] developed a target-dependent approach to TSA that isolated and evaluated sentiments regarding entities of interest. Liu et al. [2015] created a topic-adaptive model that used common and topic-adaptive features.

2.2.3. Techniques to Address Stream-Based Generation and Temporal Dependency. The volume of tweets in the Twitter stream (over 500 million tweets per day) presents significant computational challenges to performing large-scale TSA. The stream also changes rapidly, topics and novel language emerge and subside quickly, and may be unknown to TSA approaches at the time of analysis. Few researchers have addressed the challenges associated with the volume and velocity requirements of performing TSA on the Twitter stream. Tweets in the stream are also temporally dependent; a tweet is dependent upon the tweets that precede it in the stream. And a tweet may not represent a complete communication, as users commonly express a single thought through multiple sequential tweets to circumvent length restrictions.

Calibrating a machine-learned sentiment classifier typically involves passing over a training set of instances iteratively for multiple training epochs. The majority of machine-learned TSA approaches utilize this form of training, treating tweets as independent instances. However, this does not account for the underlying behavior of tweet generation, and preceding tweets in the stream. Researchers have transformed the TSA problem to more accurately reflect the stream-based generation of tweets, and devised sequence labeling approaches to sentiment classification using Markov [Vanzo et al. 2014] or perceptron [Aston et al. 2014] models. Bifet and Frank [2010] used sliding windows to subdivide the stream. Rather than classifying individual tweets, Amati et al. [2014] estimated the proportion of sentiments in the stream.

3. BENCHMARK EVALUATION OF STATE-OF-THE-ART SYSTEMS IN TWITTER SENTIMENT ANALYSIS

To assess the performance of state-of-the-art sentiment analysis systems in TSA, we conducted an extensive benchmark evaluation in tweet sentiment classification across five distinctive Twitter data sets. 28 of the top academic and commercial sentiment analysis systems and techniques were included in the evaluation, presented in Table II. They were selected from the published literature, freely available academic systems, and commercial systems requiring payment for use. While the selected systems are not an exhaustive list, they are prominent among state-of-the-art approaches to sentiment analysis. Included were well-established academic systems from the sentiment analysis literature that have demonstrated particularly strong performances in other genres of communication and are also utilized for TSA, such as the SVM Baseline [Pang et al. 2002], OpinionFinder [Riloff and Wiebe 2003], LightSIDE [Mayfield and Rose 2012], and RNTN [Socher et al. 2013]. Also selected were academic systems that have performed well in TSA in prior studies and employ some of the reviewed techniques devised to address the challenges associated with TSA, such as Sentiment140 [Go et al. 2009], SentiStrength [Thelwall et al. 2010], BPEF [Hassan et al. 2013], and FRFF [Sharif et al. 2014]. Systems from recent International Workshop on Semantic Evaluation's Sentiment Analysis in Twitter competitions (SemEval SAT) were also selected, including the winning systems from 2013 (NRC [Mohammad et al. 2013]), 2014 (TeamX [Miura et al. 2014]), and 2015 (Webis [Hagen et al. 2015]). The commercial systems selected for the evaluation were identified primarily through keyword search. Commercial systems were required to have demonstrated success in the sentiment

analysis marketplace to be included, by having hundreds of paying customers, thousands of downloads, or millions of API calls. The academic and commercial systems included in the evaluation were API-based, downloaded as desktop applications, or implemented based upon the details in the published research.

Table II. Sentiment Analysis Systems Incorporated in Benchmark Evaluation

System	**Academic / Commercial**	**General-Purpose / Domain-Specific**
AiApplied	Commercial	General-Purpose
Anonymous	Commercial	General-Purpose
BPEF [Hassan et al. 2013]	Academic	Domain-Specific
ChatterBox	Commercial	General-Purpose
EWGA [Abbasi et al. 2008]	Academic	Domain-Specific
FRFF [Sharif et al. 2014]	Academic	Domain-Specific
FRN [Abbasi et al. 2011]	Academic	Domain-Specific
GU-MLT-LT [Gunther and Furrer 2013]	Academic	Domain-Specific
Intridea	Commercial	General-Purpose
KLUE [Proisl et al. 2013]	Academic	Domain-Specific
LightSIDE [Mayfield and Rose 2012] (Version 2.0)	Academic	Domain-Specific
Lymbix	Commercial	General-Purpose
MLAnalyzer	Commercial	General-Purpose
NRC [Mohammad et al. 2013] (Version EmoLex 0.92)	Academic	Domain-Specific
OpinionFinder [Riloff and Wiebe 2003] (Version 1.5)	Academic	General-Purpose
Repustate	Commercial	General-Purpose
RNTN [Socher et al. 2013] (Version CoreNLP 3.4)	Academic	Domain-Specific
Semantria	Commercial	General-Purpose
Sentiment140 [Go et al. 2009]	Academic	General-Purpose
SentimentAnalyzer	Commercial	General-Purpose
SentiStrength [Thelwall et al. 2010] (Version .NET)	Academic	General-Purpose
SVM Baseline [Pang et al. 2002] (Version RapidMiner 5.3)	Academic	Domain-Specific
TeamX [Miura et al. 2014]	Academic	Domain-Specific
Textalytics	Commercial	General-Purpose
TextProcessing	Commercial	General-Purpose
uClassify	Commercial	General-Purpose
ViralHeat	Commercial	General-Purpose
Webis [Hagen et al. 2015] (Version SemEval 2015)	Academic	Domain-Specific

The two broad classes of sentiment analysis approaches are represented in the selected systems, those applying sentiment lexicons and scoring algorithms, and those using machine-learned models for classification. Two classes of machine-learned models are considered, systems pre-trained on a general sentiment analysis corpus requiring no additional calibration for application, and others trained within each domain of application. General-purpose sentiment analysis systems are convenient and ready to use, but lack exposure to domain-specific expressions of sentiment which may limit their performance. Domain-specific systems require large data sets with sentiment class labels to learn from before application. We next describe the systems/techniques selected for the benchmark evaluation and their approaches to sentiment analysis. We provide a description of each of the academic systems, and details on the commercial systems if available; firms developing commercial systems limit the information published on their underlying sentiment analysis approaches to protect their proprietary technology.

Among the commercial systems selected for the TSA benchmark evaluation, few offered limited information on their approaches to sentiment analysis. The Lymbix [2015] system evaluates sentiments using emotional lexicons, developed by extracting text segments from social media streams and rating them for sentiment

using a network of human raters. Intridea [2015] uses the SVM machine learning algorithm in their sentiment analysis system. The Semantria [2015] system uses search engine querying, which determines the likelihood of a phrase being used with known positive or negative terms. A number of systems utilize product review data sets for sentiment analysis training, since the text reviews are accompanied by their numeric ratings, including the TextProcessing [2015] and uClassify [2015] systems.

Among the systems based upon academic research, several are general-purpose and require no additional calibration. These systems include OpinionFinder, Sentiment140, and SentiStrength. OpinionFinder [Riloff and Wiebe 2003] uses high-precision subjectivity classifiers to identify statements containing opinions, which are then evaluated using a sentiment lexicon. The Sentiment140 system [Go et al. 2009] uses a maximum entropy-based machine-learned classifier trained on a large Twitter corpus using distant supervision. Expanding the training corpus aims to address the TSA challenge associated with sentiment class imbalance and improve the recall of sentiment classes. Sentiment140 also performs Twitter-specific preprocessing correcting for stylistic repetition of letters and replacing usernames and hyperlinks with equivalence class tokens. The SentiStrength system [Thelwall et al. 2010] also uses preprocessing, correcting spelling and slang to cope with the Twitter language, and boosts the weight of emphasized expressions before applying a sentiment lexicon.

Other academic systems selected for the benchmark evaluation are domain-specific and require training within the domain of application prior to tweet sentiment classification. These systems include the SVM baseline, entropy-weighted genetic algorithm (EWGA), feature-relation network (FRN), LightSIDE, recursive neural tensor network (RNTN), and bootstrap parametric ensemble framework (BPEF). The SVM baseline system follows established sentiment analysis approaches developed for other genres of communication [Pang et al. 2002] but widely applied for TSA. The system utilizes word n-gram features to represent tweets, feature selection using the information gain heuristic, and the SVM machine learning algorithm for classification. RapidMiner was used for the SVM implementation [Jungermann 2009]. The EWGA system utilizes the entropy-weighted genetic algorithm for selecting the features applied in the classification [Abbasi et al. 2008]. The system integrates a broad collection of syntactic and stylistic features, the EWGA for feature selection, and the SVM algorithm. The FRN system utilizes feature-relation networks for feature selection [Abbasi 2010; Abbasi et al. 2011]. The FRN is a rule-based text feature selection method that considers the semantic and syntactic relationships between n-grams, to efficiently remove redundant and irrelevant features from the representation. The system then applies the selected features to the SVM algorithm for sentiment classification. LightSIDE is a system for text analysis and assessment based upon semantic, syntactic, and stylistic features including word and part-of-speech grams [Mayfield and Rose 2012]. Several classifiers are available within the system, sourced from the Weka data mining package [Witten and Frank 2005]; for the benchmark evaluation, SVM was utilized. The RNTN system represents text with word vectors and a parse tree, extracted using a tensor-based compositional model [Socher et al. 2013]. The RNTN utilizes a softmax classifier to label all word and phrase vectors in the parse tree by computing the posterior probabilities of the sentiment classes for each word vector. The BPEF system employs a search framework to identify an effective classifier ensemble for sentiment classification [Hassan et al. 2013]. The system parameterizes the data sets for training, features to represent the text, and classification algorithms, to address the TSA challenges of tweet brevity and sentiment class imbalance. This approach embodies concepts

similar to the feature representation expansion and multiple classifier techniques, but extends far beyond these as the expansion, search, and classifier ensemble composition occur across the entire learning stack. BPEF uses text summarization techniques, word and part-of-speech grams and semantic features, and seven machine learning classification algorithms.

Several top-performing systems from recent SemEval SAT competitions were also included in the benchmark evaluation. GU-ML-TLT [Gunther and Furrer 2013] was the second-ranked system in the 2013 SemEval SAT competition. The system uses a stochastic gradient descent classifier coupled with a feature set comprised of normalized unigrams, stems, semantic clusters, SentiWordNet [Baccianella et al. 2010] assessments of individual words, and negation measures. KLUE [Proisl et al. 2013], which ranked fifth in the same competition, employs a maximum entropy classifier and a feature set consisting of unigrams, bigrams, tweet length, sentiment, emotion, colloquial lexicons, and negation measures. The top-ranked system in the 2013 SemEval SAT competition was the NRC system [Mohammad et al. 2013]. The system uses a rich representation encompassing over 300,000 features, including unigrams to four-word-grams, part-of-speech tags, various sentiment lexicons, emotion lexicons, punctuation marks, emoticons, word length and capitalization measures, semantic clusters, and negation measures, coupled with a linear-kernel SVM. The top-ranked systems from the more recent 2014 (TeamX [Miura et al. 2014]) and 2015 (Webis [Hagen et al. 2015]) SemEval SAT competitions were also included. The TeamX system uses two part-of-speech taggers designed for formal (Stanford) and informal (CMU ARK) texts, and incorporates unigrams to 4-grams and several sentiment, emotion, and colloquial lexicons. These features are used to train a logistic regression classifier. The Webis system uses a simple voting ensemble comprised of the NRC, TeamX, KLUE, and GU-ML-TLT system's sentiment classification probabilities. The system's performance underscored the power of ensemble methods that are comprised of a diverse set of underlying approaches.

3.1 Description of Data and Experiments

To comprehensively evaluate the systems across a variety of TSA applications, data sets from five distinctive topical domains were included in the benchmark evaluation. The selected domains were pharmaceuticals, retail, security, technology, and telecommunications. Each domain is characterized by a distinctive pattern of sentiment expression, and represents a valuable application area for TSA systems. The retail, tech, and telco domains focus on TSA for the evaluation of consumer sentiments regarding products and services, with specific implications for marketing analytics. The security and pharma domains are applications related to monitoring for security incidents or outbreaks of adverse medical events, respectively, with implications for security informatics and smart health.

The tweets in each of the data sets were evaluated by human annotators via Amazon Mechanical Turk (AMT), and labeled at the tweet level for sentiment class (positive, negative, or neutral). The technology data set was developed by Sanders [2011]. For the remaining four domains, best practices for data annotation were followed [Callison-Burch and Dredze 2010]. Prior to annotation, manual and automated preprocessing procedures were employed to remove irrelevant tweets (non-English, spam, or unrelated to the topic of interest). Within AMT, the Sentiment Rating module was utilized, with five experienced turks classifying each tweet according to the sentiment expressed in the tweet. Each turk received detailed directions for annotation, following the suggestions of Callison-Burch and Dredze

[2010]. Only the tweets with clearly expressed sentiment directed toward the topic of interest were classified as positive or negative. For example consider a positive tweet from the telco domain, regarding the topic of interest the firm Telus: 'New telus calling rates are awesome! @ammarhassan'. All other tweets were classified as neutral. Tweets with sentiment expression unrelated to the topic of interest were classified as neutral. For example, 'Thank you to @GSMLiberty for unlocking my phone for 1/5 the cost of what @TELUS was asking for! Keep up the good work!'. Tweets with mixed sentiments (expressing both positive and negative sentiments in the tweet) were also classified as neutral. For example, 'If you have #DeviceProtection you're safe! RT @mdocc: i hope telus will fix my phone even tho it has water damage..'. Objective tweets or tweets posing questions related to the topic of interest without the expression of sentiment were classified as neutral. For example, 'Shall I get it? Hmmm... RT @TELUS: Instagram - Fast beautiful photo sharing now available for Android'. Unanimous agreement among the five human annotators regarding the sentiment classification of a tweet was required for the tweet to be retained in a data set for the benchmark evaluation. Tweets disagreed upon by the human annotators were removed from the data sets.

Descriptions of the five data sets included in the benchmark evaluation are presented in Table III. Each data set consisted of several thousand tweets, with three of the five having more than 5,000. The majority of tweets from each domain expressed neutral sentiment, which confirmed our expectations regarding the infrequent sentiment expressions in Twitter. The exception was the retail domain, which contained frequent expressions of positive sentiments regarding consumer experiences with products. Each data set had heavily imbalanced sentiment classes.

For the benchmark evaluation, general-purpose systems classified the tweets in each of the five data sets, while domain-specific systems performed 10-fold cross validation within each domain. Several standard evaluation metrics were considered, overall accuracy, and class-level recall and precision. Overall accuracy is the percentage of tweets classified correctly (as positive/negative/neutral). Class-level recall is the percentage of tweets from a given class that were correctly identified as belonging to that class. Class-level precision is the percentage of tweets accurately classified as belonging to a given class. Due to space constraints we present a portion of the benchmark evaluation results. The overall accuracies for all systems are reported, and class-level recall rates for select systems.

Table III. Description of Twitter Data Sets used in Evaluation

Data Set	Description of Tweets	Total Tweets	Polarity Class Distribution		
			Positive	Negative	Neutral
Pharma	Related to users' experiences with pharmaceutical drugs. Includes mentions of adverse events and positive interactions.	5,009	15.6%	11.1%	73.3%
Retail	Includes discussion of a category of retail products (household paint) and user experiences related to those products.	3,750	42.7%	9.0%	48.3%
Security	Related to major security companies' products and services, including security incidents and new software releases.	5,086	24%	11.1%	64.9%
Tech	Related to four major tech firms. Includes discussion of products, services, policies, and user experiences.	3,502	15.1%	16.9%	68.0%
Telco	Related to telecommunications company Telus' products and services. Includes discussion of experiences, news, and events.	5,281	20.9%	8.9%	70.2%

3.2 Benchmark Evaluation Results

The results of the benchmark evaluation in tweet sentiment classification are presented in Table IV for each system and domain. Also reported is the average accuracy across domains, as an overall indicator of performance. On the whole, the systems performed poorly with a wide range of average classification accuracies from 40% to 71%. Domain-specific approaches to sentiment analysis widely outperformed the general-purpose approaches. Although they required training data with sentiment class labels, they were able to capture domain-specific indicators of sentiment expression which improved their performance. The average results across all general-purpose and all domain-specific systems are also presented in Table IV.

Table IV. Benchmark Evaluation Results - System Classification Accuracy

System	Average	Pharma	Retail	Security	Tech	Telco
AiApplied	61.84	69.59	47.99	64.05	60.39	67.20
Anonymous	40.86	33.65	49.93	32.71	43.11	44.89
BPEF	71.38	67.81	65.24	75.32	76.30	72.21
ChatterBox	67.43	75.04	53.19	67.20	69.73	71.99
EWGA	68.12	70.21	60.00	68.50	70.50	71.41
FRFF	70.72	62.86	68.76	73.97	74.90	73.11
FRN	69.17	72.60	59.96	69.98	71.00	72.30
GU-MLT-LT	60.60	45.32	68.21	57.81	60.25	71.41
Intridea	63.31	64.18	47.37	62.63	75.19	67.20
KLUE	62.78	55.60	71.15	54.27	62.25	70.65
LightSIDE	69.35	70.71	58.22	69.86	76.99	70.99
Lymbix	56.63	52.03	54.81	47.60	63.45	65.25
MLAnalyzer	45.20	37.95	52.15	41.35	48.06	46.47
NRC	71.33	75.26	64.93	76.39	64.96	75.08
OpinionFinder	57.66	57.08	52.40	55.01	56.94	66.86
Repustate	43.98	35.80	41.06	31.93	40.90	70.20
RNTN	61.47	66.76	55.25	64.69	55.51	65.14
Semantria	53.50	44.68	56.33	45.46	60.99	60.06
Sentiment140	66.46	62.09	61.77	68.84	67.82	71.79
SentimentAnalyzer	55.15	55.33	51.36	54.83	56.50	57.75
SentiStrength	67.49	74.68	56.35	65.51	69.61	71.31
SVM Baseline	66.86	67.50	59.52	66.02	70.02	71.22
TeamX	67.20	57.60	70.35	62.82	69.10	76.14
Textalytics	66.22	70.33	55.14	66.33	68.29	71.02
TextProcessing	54.06	49.68	50.01	58.40	52.40	59.79
uClassify	47.22	51.70	42.12	47.51	50.31	44.47
ViralHeat	61.16	63.77	48.42	61.94	64.12	67.56
Webis	71.41	76.16	64.40	77.37	63.68	75.46
All General-Purpose Systems	56.76	56.10	51.28	54.46	59.24	62.74
All Domain-Specific Systems	67.53	65.70	63.83	68.08	67.96	72.09

Among the general-purpose systems, ChatterBox, Sentiment140, SentiStrength, and Textalytics performed best, with average sentiment classification accuracies above 66%. ChatterBox generated the best classification performance on average, but Sentiment140's performance was most consistent across the five domains, with accuracies ranging from 60% to 71%. A number of general-purpose systems performed very poorly, with four systems producing overall accuracies below 50%. The average classification accuracy across general-purpose systems was only 56%.

Performances among the domain-specific systems were better, with average classification accuracies ranging from 61% to 71%. Overall, the domain-specific

systems averaged 67%, an 11% improvement on that of the general-purpose systems. The top domain-specific systems were Webis, NRC, FRN, FRFF, LightSIDE, and BPEF, each producing average classification accuracy over 69%. Three of these systems also surpassed 71% on average and were consistent performers across the five domains. Overall, we consider BPEF, NRC, and Webis as the best systems in the benchmark evaluation in terms of classification accuracy.

The sentiment class-level recall rates from the benchmark evaluation are also presented in Table V for select systems. It is clear based upon the results that class-level recall rates vary considerably across sentiment classes for most systems. For example, the SentiStrength system achieved good classification accuracy overall due in part to high negative class recall in the pharma and security data sets (90.47 and 90.93, respectively), while its neutral class recall rates in these domains were 29.29 and 0.15, respectively. Such imbalances in class-level recall indicate bias toward a particular sentiment class or classes. Similarly, ChatterBox tended to recall positive and neutral tweets but had much lower negative class recall rates. Textalytics exhibited higher neutral class recall rates but much lower positive and negative recall, indicating a lack of sensitivity in the sentiment analysis.

Table V. Benchmark Evaluation Results – Select System's Sentiment Class-level Recall

System	Pharma			Retail			Security			Tech			Telco		
	Pos	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**	**Pos**	**Neg**	**Neu**
BPEF	63.2	61.3	69.8	60.7	64.6	69.3	75.9	72.4	75.6	69.3	78.6	77.3	57.0	74.5	76.5
ChatterBox	37.2	56.7	62.5	57.6	11.0	52.8	63.4	30.1	66.5	52.6	45.1	56.5	53.2	34.0	64.4
FRN	39.2	21.0	84.9	59.0	24.7	67.4	55.2	35.9	81.2	37.2	42.1	80.9	33.3	37.3	81.1
Intridea	26.2	68.7	37.5	32.5	63.7	37.3	32.9	62.2	39.9	69.8	81.8	74.8	32.7	68.4	46.3
LightSIDE	44.9	28.1	82.6	57.2	34.2	63.5	56.6	47.3	78.6	47.9	55.9	84.3	38.7	49.8	83.3
NRC	60.3	26.1	85.9	74.3	43.2	60.7	62.2	46.3	86.7	42.5	35.5	77.3	49.3	56.6	85.1
OpinionFinder	32.2	55.9	62.5	20.9	18.2	86.6	33.9	42.2	65.0	14.0	13.7	77.3	23.2	42.3	83.0
RNTN	32.5	16.7	81.6	39.5	14.9	76.7	43.1	14.8	81.2	16.4	13.2	74.8	35.3	20.3	79.7
Sentiment140	44.0	62.6	65.8	43.8	11.0	87.1	61.0	26.0	79.0	55.1	36.8	78.4	46.4	28.4	84.8
SentiStrength	47.0	90.5	29.3	53.3	38.1	62.4	87.9	90.9	0.2	62.0	45.3	61.9	59.7	80.6	42.5
Textalytics	28.8	23.7	86.2	21.0	10.4	93.6	24.8	16.0	90.3	27.0	18.0	90.0	25.1	29.1	90.0
Webis	58.4	28.8	87.1	77.7	37.2	57.7	66.7	50.7	85.8	46.2	37.2	74.2	51.5	56.6	85.0

The issue of bias toward a particular sentiment class, evidenced in imbalanced class-level recall rates, impacted not only general-purpose systems but domain-specific as well. The LightSIDE, FRN, and RNTN each had much higher neutral class recall rates than positive or negative classes, demonstrating a limited ability to detect sentiment expressions. The exception was the BPEF system, which performed with relative consistency in recalling each sentiment class. The BPEF positive, negative, and neutral class-level recall rates were within 10% of one another in four of the five domains. Additionally, the system had the best overall positive recall rate in two of the five domains, and negative recall rate in one domain. Sentiment class-level recall rates have important implications in TSA applications. When applied in prediction or detection, TSA systems with lower positive and negative recall rates generate time series indices with less variation, which are less effective in detecting events characterized by fluctuations in Twitter sentiment [Hassan et al. 2013].

3.3 Benchmark Evaluation Error Analysis

As shown in the benchmark evaluation results, the performances of the systems were lackluster overall, with an average tweet sentiment classification accuracy of 61% across all systems and domains. The best systems achieved only 71% accuracy. There was also a wide range in the average classification accuracies across systems (31%).

System performances varied considerably across domains, and suffered from imbalanced class-level recall rates and a bias toward neutral classification.

The poor performances of state-of-the-art systems in the benchmark evaluation underscore the challenges and complexities associated with TSA. To better understand the root causes of these poor performances, and uncover why errors in sentiment classification occurred to improve the next generation of TSA approaches, a detailed error analysis was performed. The 1,000 most misclassified tweets in each of the five domains were selected for the error analysis. Following the best practices outlined in prior studies [Wiebe et al. 2005], three human evaluators analyzed each of the misclassified tweets to determine why a TSA approach may have mistakenly interpreted the sentiment expressed. As part of their evaluation process, the evaluators began by using Appraisal Theory [Scherer 1999; Martin and White 2005], and its related literature from the natural language processing community, to derive an initial set of potential categories for classification error [Whitelaw et al. 2005]. With its roots in psychology, Appraisal Theory suggests that sentiments and emotions are derived from our evaluations (or appraisals) of events, individuals, or situations. These appraisals vary for different individuals, and the resulting sentiments and emotions triggered, and how these manifest, also vary [Scherer 1999]. Variations in appraisals may result in sentiment classification error, as the sentiments expressed by an author in a tweet are evaluated differently by annotators and classifiers. Some prominent dimensions of variation have been identified, including the force/intensity of sentiment (e.g., strong or subtle), focus of sentiment on a particular target (i.e., relevance to the appraised), inclusion of multiple targets, use of humor or other literary devices, appraisal intention (e.g., promotion), and idiosyncratic expression mechanisms. Some prior work has attempted to categorize variations in how sentiments are expressed into appraisal groups [Whitelaw et al. 2005]. Using such an approach, and leveraging prior literature, our evaluators derived a preliminary set of categories of tweet sentiment classification error. Examples included “subtle positive” and “subtle negative” sentiment. This preliminary set of categories developed using Appraisal Theory was refined and supplemented through multiple rounds of analysis of the misclassified tweets and discussion amongst the evaluators. After much consideration by the evaluators, a taxonomy of causes for tweet sentiment classification error was finalized. The misclassified tweets were then categorized according to the taxonomy. Errors attributed to multiple classes were assigned to the most salient category.

The taxonomy of tweet sentiment classification errors consisted of 13 categories. A pie-chart cloud representing the taxonomy is presented in Figure 1. Each pie-chart represents a category in the taxonomy and possible cause for classification error. The size of a pie corresponds to the relative frequency of misclassification associated with the category across domains, and the percentages of total misclassifications are presented. Within each pie-chart, the distribution of misclassifications associated with the category across the five domains is shown. The 'neutral mistaken for sentiment' category includes neutral-sentiment questions or requests mistaken for positive compliments or negative criticisms. For example, a tweet containing "it would be great if we could" would be misclassified as positive. The 'humor' category includes jokes, sarcasm, rhetoric, and related devices cited in prior studies as problematic. The 'marketing' category includes tweets describing events, contests, and advertisements, which are considered by the human annotators to be neutral but often classified as positive. The 'atypical usage' category includes misclassifications attributed to an alternative usage of terms. For example, a harsh curse word used to

express an extremely positive sentiment. The 'subtle positive sentiment' category contained misclassified tweets with subtle sentiment cues related to donations, charities, or events with positive connotation. These tweets were often classified as positive by the human annotators, while TSA approaches lacking information on the subtle sentiment terms mistakenly classify them as neutral. A common error in each domain was misclassifications due to the expression of sentiments irrelevant to the topic of interest. The 'mixed sentiments' category referred to tweets expressing both positive and negative sentiments toward a topic.

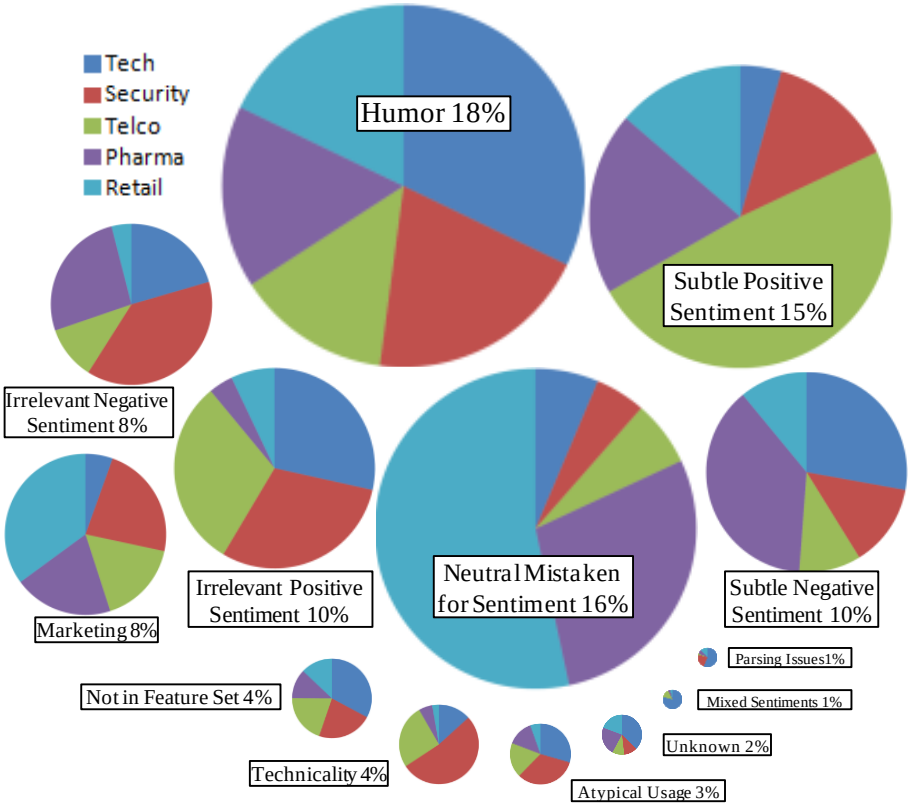

Fig. 1. Taxonomy of Sentiment Classification Errors.

Using the sentiment classification error taxonomy, the frequently occurring errors in each of the five domains were examined. The distributions of errors are presented in Figure 2. Misclassifications due to 'humor' were common, accounting for 10% to 15% of total error in each domain. This finding was somewhat expected given the known difficulties sentiment analysis approaches have in evaluating jokes, sarcasm, rhetoric, and related literary devices. The inability to detect the subtle sentiment cues expressed in the 'subtle positive sentiment' category resulted in over 10% of the errors in four of the five domains. Sentiments irrelevant to the topic of interest (either positive or negative) accounted for more than 10% of the misclassifications in multiple domains. However, errors related to the expression of mixed sentiments about a topic were relatively infrequent, despite posing a major problem in other genres of communication like web forums or blogs. The length limitation imposed on tweets inhibits the expression of multiple sentiments in such brief communications.

Specific categories of error were found to be most prevalent in particular domains, and the errors in a given domain were often related to only a few categories. 'Neutral

mistaken for sentiment' was a significant cause for misclassification in pharma and retail domains, accounting for 23% and 42% of error respectively. In these domains, questions from consumers regarding patient experiences with prescription drugs or the quality of household paints were mistakenly classified as expressing negative sentiments. The 'subtle positive sentiment' category accounted for 38% of error in the telco domain. Many tweets in this data set discussed charitable activities, donation drives, and fund-raising events, misclassified as expressing neutral sentiment. While the 'humor' category was responsible for a relatively large portion of error in each domain, errors of this category were prevalent in the tech and security domains. Tweets from these domains often featured jokes, sarcasm, and rhetoric regarding security software, smart phone makers, or major technology firms, which were difficult to classify.

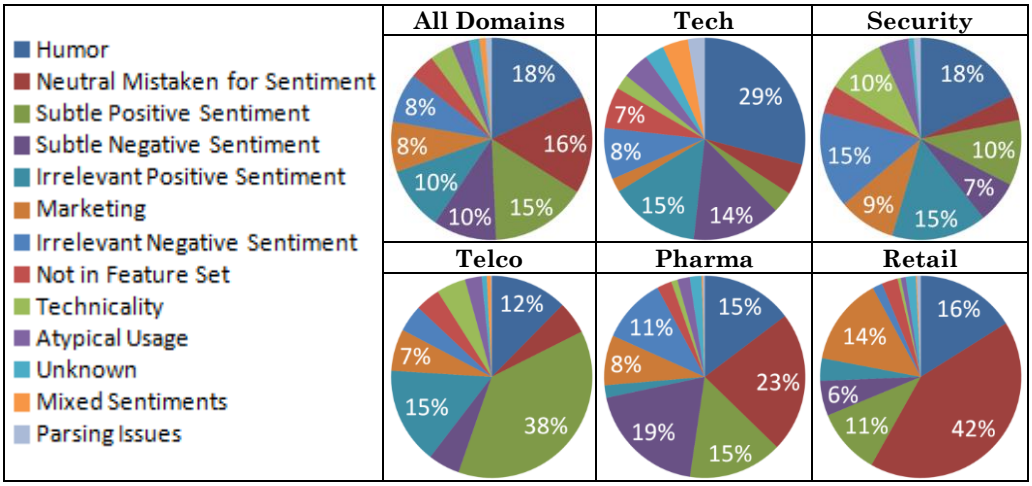

Fig. 2. Distribution of Classification Errors across Domains.

Overall, the error analysis provided insights into the possible causes of tweet sentiment misclassifications across domains, and considerations for improvements in the next generation of TSA approaches. Several categories of error indicated the need for increased adoption of the techniques devised to address the challenges associated with TSA. For example, the systems that misclassified tweets due to errors in the 'humor' category may benefit from incorporating Twitter-specific features into their analysis related to jokes, sarcasm, and other humorous rhetoric devices. Classification errors associated with 'subtle positive sentiment' or 'subtle negative sentiment' may be reduced if TSA approaches utilized the sentiment information propagation technique to associate known sentiment expression with these subtle cues. The multiple classifier technique may improve sentiment analysis in cases of tweets with neutral statements mistaken for conveying sentiment, through the application of a subjectivity classifier to identify questions or requests, followed by a sentiment classifier to evaluate subjective tweets.

The results of the error analysis also emphasized the importance of careful annotation of tweets to train and evaluate the sentiment analysis systems. The human annotators considered 'marketing' tweets related to events, contests, and advertisements as neutral, when they may have reasonably classified these as positive given a different interpretation or additional context, reducing the number of misclassifications associated with this category. The annotators were also directed to only classify tweets with clearly-expressed sentiment directed toward the topic of

interest as positive or negative. As a result, systems detected sentiments directed to other topics and misclassified these tweets as positive or negative. Assigning tweets to a single sentiment class when multiple sentiments directed to the topic of interest are expressed is a challenging annotation requiring careful scrutiny. Although great effort was expended to ensure the quality and accuracy of the annotated Twitter data, additional human annotators or more-detailed guidance may have mitigated the effects of these errors on the systems in the benchmark evaluation.

3.4 Benchmark Evaluation Event Detection Case Study

TSA systems are often deployed to monitor Twitter and detect the occurrences of specific events. To further the benchmark evaluation we applied the top-performing systems in an event detection case study. While the benchmark evaluation in tweet sentiment classification assessed the systems in classification performance, the case study evaluates how these performances may translate to an operational, application context. The case focuses on Telus, a major Canadian telecommunications firm. The period of analysis spanned five years, including over 150,000 tweets posted regarding Telus. Two independent annotators scrutinized the tweets and identified 20 events that generated significant sentiment expressions, 11 positive and 9 negative events.

Several systems that performed in the top ten on our benchmark telco data set were selected for the case, including Webis, BPEF, FRN, SentiStrength, and Sentiment140. Each of these methods produced sentiment classification accuracy of at least 71% on the telco data set. To better understand the interplay between classification performance and performance in TSA application contexts, we also included several systems that performed in the bottom ten: uClassify, Anonymous, MLAnalyzer, SentimentAnalyzer, and TextProcessing. Each of these systems produced sentiment classification accuracies below 60% on the telco data set.

Domain-specific systems were first trained on the benchmark telco data set, then applied to classify the 150,000+ tweets related to Telus. The sentiment classifications produced by each system were then utilized to construct average monthly sentiment time-series. Following the approach adopted in prior TSA event detection studies [Sharif et al. 2014; Abbasi and Adjeroh 2014], we employed a control chart method. For each system's generated time-series, if the sentiment value in the month of an event was one standard deviation above (for positive events) or below (for negative events) the mean, the event was considered to be detected correctly by the system.

The event detection evaluation results are presented in Table VI. The first three columns depict the overall detection rates across the 20 events, and the breakdown across the 11 positive and 9 negative events. For reference and comparison, the remaining columns present the previously discussed benchmark results on the telco data set (classification accuracy and class-level recall rates). A few interesting observations emerge. First, while sentiment classification accuracies on the telco data set ranged from 44% to 75%, performances in event detection varied more widely from 5% to 65%. This suggests that classification performance has a major impact on TSA applications, where these classifications are ultimately utilized. Second, as expected, the sentiment classification performance of TSA methods is correlated with their effectiveness in TSA application. The five systems with low classification accuracy detected 15% or less of the 20 events, suggesting that poor classification performance results in time-series signals with too much noise. Third, among top-performing systems where sentiment classification accuracies were relatively comparable (71% to 75%), the percentage of events detected varied

substantially from 30% to 65%. This suggests that differences in class-level recall rates may impact system performances in TSA event detection applications.

Table VI. Benchmark Evaluation Results – System Event Detection Performance

System Type	System	Event Detection Performance			Benchmark Sentiment Classification Performance			
		Overall	Positive Events	Negative Events	Accuracy	Positive Recall	Negative Recall	Neutral Recall
High sentiment polarity classification accuracy	BPEF	65.0	54.5	77.8	72.2	57.0	74.5	76.5
	Webis	55.0	72.7	33.3	75.5	51.5	56.6	85.0
	FRN	40.0	54.5	22.2	72.3	33.3	37.3	81.1
	Sentiment140	40.0	27.3	55.6	71.8	46.4	28.4	84.8
	SentiStrength	30.0	45.5	11.1	71.3	59.7	80.6	42.5
Low sentiment polarity classification accuracy	TextProcessing	15.0	18.2	11.1	59.8	41.2	50.2	66.6
	SentimentAnalyzer	10.0	9.1	11.1	57.8	45.4	48.9	62.7
	MLAnalyzer	5.0	9.1	0.0	46.5	70.8	57.3	37.9
	Anonymous	5.0	9.1	0.0	44.9	74.7	71.8	32.6
	uClassify	5.0	9.1	0.0	44.5	44.2	23.7	42.7

To further investigate this last point, Figure 3 depicts the time-series for four of the top-performing methods. The x-axis of the graph depicts time in months while the y-axis shows sentiment ranging from -1 (extremely negative) to 1 (extremely positive). Several of the aforementioned significant events regarding Telus that occurred during the period of analysis were also noted on the timeline. As shown in the figure, the sentiment time-series indices constructed using the classifications generated by the systems varied considerably, with positive and negative sentiment spikes. Many of these fluctuations coincide with specific events regarding Telus. For example, when it was revealed that Telus jokingly referred to their customers as deadbeats in their terms-of-use, the incident went viral with strong negative sentiment. This

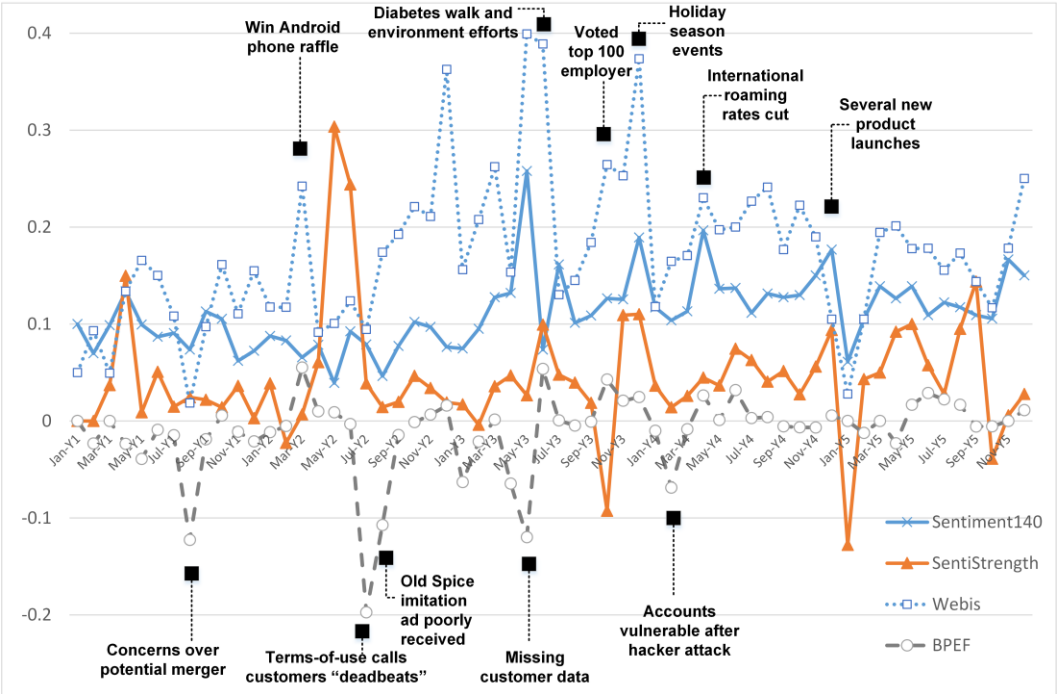

Fig. 3. Sentiment Time-Series Indices Generated by Top Twitter Sentiment Analysis Systems for Telus Telco Tweets.

reaction was reflected in a sharp drop in the BPEF sentiment index. Similarly, positive opinions in response to an Android phone raffle were indicated by the BPEF and Webis systems through pronounced increases in Twitter sentiment. All four systems effectively identified positively-received Telus events associated with the holiday season in December of year 3. Overall, BPEF and Webis were more responsive to the events (with event detection rates greater than 50%). As previously alluded to, this was attributable to differences in positive, negative, and neutral recall rates; relative to other high-accuracy methods, BPEF was the most balanced in its class-level recall rates – the system's greater sensitivity to positive and negative cues and balanced recall across sentiment classes produced a rich representation of Twitter sentiments and improved ability to detect the occurrences of important Telus events. The case study sheds light on the interplay between TSA methods' classification performance and TSA applications. The broader implications are elaborated upon in the ensuing section.

4. KEY TRENDS AND TAKEAWAYS

The results of our benchmark evaluation, error analysis, and application case study have several important takeaways for researchers developing TSA techniques or applying TSA for various social media analytics use cases.

4.1 Importance of Using/Developing Systems that Support Domain Adaptation

Our benchmark evaluation included 12 systems incorporating supervised machine learning methods that could be easily trained on data from the application domain, and 16 general-purpose systems relying on pre-defined rule sets. As previously discussed in the benchmark evaluations section and Table IV, on average, the domain-specific systems outperformed their general-purpose counterparts by 11 percentage points in terms of overall classification accuracy. This performance delta was attributable to the domain-specific systems' better detection of tweets containing positive or negative sentiment polarity. Figure 4 presents the mean positive recall (left chart) and negative recall (right chart) across all general-purpose and domain-specific systems for the five benchmark data sets. For most data sets, the domain-specific systems had positive/negative recall rates that were 12% to 15% higher than those attained by the general-purpose systems. Simply put, the general-specific systems' rule sets failed to include many important sentiment polarity cues. For instance, in the telco context, statements such as "leaving" or "switching from" signify strong negative polarity (i.e. customer churn), which the domain-specific systems were better able to detect and interpret.

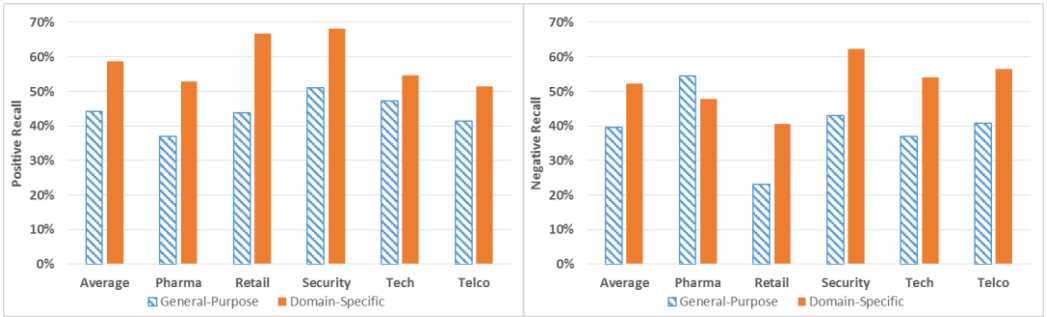

Fig. 4. Positive and Negative Recall Rates for General-Purpose and Domain-Specific Systems across Data Sets in Benchmark Evaluation.

4.2 Effectiveness of Ensemble Methods

Three of the top four performers in our benchmark evaluation used an ensemble of machine learning classifiers. Webis [Hagen et al. 2015] utilizes an ensemble comprising of four existing systems also incorporated in our evaluation: NRC, KLUE, TeamX, and GUMLTLT. BPEF incorporates a parametric ensemble featuring combinations of different classifiers, feature sets, and reference data sets [Hassan et al. 2013]. FRFF employs an ensemble of SVM models trained using different feature set combinations [Sharif et al. 2014]. All three of these methods attained average accuracies of over 70% across the benchmark data sets, even outperforming methods that included state-of-the-art deep learning methods [e.g., Socher et al. 2013]. BPEF and Webis were also the top two performing systems in the event detection case study. Ensemble methods seem well-suited for overcoming the class imbalance and poor sentiment recall challenges plaguing many existing TSA techniques.

4.3 Importance of Including an Array of Lexicons and Linguistic Resources

We discussed the importance of learning domain-specific sentiment cues. However, the top-performing systems balanced domain adaptation/learning with extensive use of manually crafted and automatically constructed general-purpose lexicons and other linguistic resources. Figure 5 shows the lexicon/resource usage frequency breakdown across the top-four performing systems (Webis, NRC, BPEF, and FRFF). Collectively, these 4 systems used 30 lexicons/resources, including SentiWordNet, the MPQA, Bing Liu, and Sentiment140 lexicons, AffectWordNet, WordNet, other emotion lexicons, named entity lexicons, negation/boosting lexicons, Twitter POS taggers, and other Twitter-specific hashtag and emoticon lexicons. In order to offset the brevity and novel language usage challenges presented by Twitter, the extensive use of lexicons and linguistic resources seems essential for supporting the feature representation expansion; (2) Twitter-specific preprocessing; (3) and feature construction previously discussed in Table I and section 2.2.1.

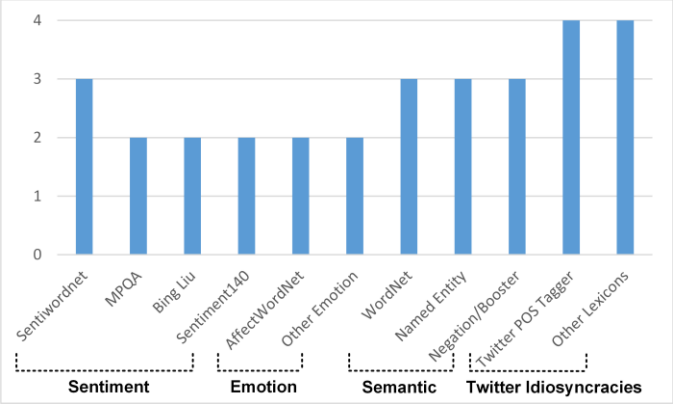

Fig. 5. Frequency of Lexicon and Other Linguistic Resource Usage by Top-four Performing Systems.

In particular, we believe TSA systems should incorporate lexicons and linguistic resources that are grounded in relevant theories from the psychology, language, and communications literature that have been incorporated in natural language processing research. One such example, mentioned earlier in our discussion of the error analysis taxonomy, is Appraisal Theory. TSA systems would undoubtedly benefit from inclusion of features that make rigorous affordances for the array of manners in which sentiments are expressed [Whitelaw et al. 2005]. Other example

theories/frameworks include System Functional Linguistic Theory [Halliday 2004; Abbasi and Chen 2008], Language Action Perspective [Winograd 1986; Abbasi et al. 2018], and the Geneva Emotion Wheel [Scherer 2005].

4.4 Error Analysis Reveals Several Existing Challenges Facing TSA Techniques

The error analysis revealed that several challenges remain. Existing TSA techniques continue to misinterpret user intentions and purpose, particularly regarding suggestions and questions. Complex literary devices such as sarcasm and rhetoric also remain problematic. Marketing and promotion verbiage is often mistakenly interpreted as positive. Parsing issues, use of nuanced sentiment cues not in existing lexicons or training data, and tweets containing mixed emotions/opinions also pose problems. These results suggest that mechanisms for handling sentiment analysis tasks such as sentiment target detection, phrase/aspect-level sentiment analysis, detection of literary devices (e.g., sarcasm) need to be better integrated into the core polarity detection engines. Fortunately, over 95% of errors were attributable to a potentially addressable category within our taxonomy – most errors are associated with areas where research is in progress across the broader CS and IS communities.

4.5 TSA Technique Performance has Serious Implications for Various Application Areas

Our event detection case study shed light on the interplay between TSA techniques' sentiment classification performance and its implications for TSA applications. The case study revealed that not only are application results (in this case event detection) correlated with underlying technique-classification performance, but class-level recall rate balance/imbalance can also impact detection rates for positive and negative events. Studies incorporating social media sentiment variables as input for various applications tasks such as financial forecasting, health surveillance, election outcome prediction, adverse event detection, etc. should carefully evaluate the TSA techniques incorporated and report these results. For instance, one could easily use an inferior TSA technique and infer that Twitter sentiments are not meaningful in a given application. Hence, checks to ensure validity of the sentiment constructs incorporated in social media studies are essential. Additionally, our error analysis results suggest that there is potential for "gaming" social media monitoring systems: inflating sentiments via gimmicky tweets, or deflating through spammy negative content.

5. CONCLUSION

Twitter is a major social media platform that has experienced tremendous growth in communication volume and user membership worldwide. Many researchers and firms have recognized that valuable insights on issues related to business and society may be achieved by analyzing the opinions expressed in the abundance of tweets. However, the clarity of these insights and the effectiveness of the derived sentiment information when applied are critically dependent upon the underlying TSA approach and its ability to accurately evaluate the opinions expressed by users. State-of-the-art TSA approaches continue to perform poorly, with reported sentiment classification accuracies typically below 70%. Considering the popularity of Twitter, value of the information derived through TSA, and difficulties experienced by state-of-the-art TSA approaches, a thorough investigation of these issues was conducted.

With respect to the specific research questions examined in this study, our review of the literature revealed two major motivations for TSA research. The first focuses on the application of TSA to gain insights into various business or social issues, predict key indicators, or monitor Twitter for emerging information or events.

Recognizing the value of information derived through accurate TSA, the second focuses on innovating and developing improved techniques and approaches to TSA. Several characteristics of tweets challenge even state-of-the-art TSA approaches, including the brevity of tweets and resulting compact, novel language with Twitter-specific communication elements, a strong sentiment class imbalance, and stream-based tweet generation. Attention to these three challenging characteristics is demanded to achieve accurate TSA, and generate the benefits associated with effective application of the derived sentiment information. Through our review of the literature, we developed a taxonomy of the various techniques devised to address these challenges. Sentiment information propagation, feature representation expansion, Twitter-specific preprocessing, and Twitter-specific features addressed the challenge of tweet brevity and novel Twitter language, training set expansion, multiple classifier methods, and sentiment-topic models addressed the challenge of sentiment class imbalance and poor sentiment recall, and stream-based classifiers addressed the challenge of stream-based tweet generation and temporal dependency.

To assess the state-of-the-art in TSA, we conducted a benchmark evaluation of 28 top academic and commercial systems in tweet sentiment classification across five distinctive Twitter data sets. The results revealed the performances of the systems were lackluster overall, with an overall average sentiment classification accuracy of 61% across systems and domains. There was a wide range in the accuracies of systems (31%). Domain-specific approaches outperformed general-purpose by an average of 11%. Although they required training data with sentiment class labels, they were able to capture domain-specific indicators of sentiment expression which improved their performance. In general, system performances varied substantially across domains, and suffered from an inability to detect sentiment expressions, evidenced by imbalanced class-level recall rates. BPEF, NRC, and Webis were the best systems in terms of classification accuracy, with over 71% accuracy across the domains. However, the BPEF system outperformed NRC and Webis in terms of recall, with generally higher rates and greater consistency across sentiment classes.

To uncover the causes of commonly occurring tweet sentiment classification errors, we performed an error analysis following the experimentation. A taxonomy was developed consisting of 13 categories representing the probable causes for classification error. We described these error categories in detail, and examined their distributions across the Twitter domains, to gain insights and improve the next generation of TSA approaches. Specific categories of error were prevalent in particular domains, and errors in a domain were often related to a few categories.

Since TSA systems are often deployed to monitor Twitter and detect the occurrences of specific events, to further the benchmark evaluation we applied select systems in an event detection case study. In general, the sentiment indices constructed using the tweet classifications generated by the systems exhibited little variation over time. However, systems with greater recall for the sentiment classes provided an improved ability to detect the occurrences of significant events.

ACKNOWLEDGMENTS

The authors would like to thank collaborators in the telecommunications, health, security, tech, and retail industries for their invaluable assistance and feedback on the benchmark evaluation, error analysis, and event detection case study.

REFERENCES

A. Abbasi, "Intelligent Feature Selection for Opinion Classification," IEEE Intelligent Systems, vol. 25, no. 4, pp. 75-79, 2010.

A. Abbasi and D. Adjeroh, "Social Media Analytics for Smart Health," IEEE Intelligent Systems, vol. 29,

no. 2, pp. 60-80, 2014.
A. Abbasi and H. Chen, “CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated-Communication,” MIS Quarterly, vol. 32, no. 4, pp. 811-837, 2008.
A. Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," ACM Trans. on Information Systems, vol. 26, no. 3, 2008.
A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting Attributes for Sentiment Classification using Feature Relation Networks," IEEE Trans. on Knowledge and Data Engineering, vol. 23, no. 3, pp. 447-462, 2011.
A. Abbasi, T. Fu, D. Zeng, and D. Adjeroh, "Crawling Credible Online Medical Sentiments for Social Intelligence," Proc. of ASE / IEEE Intl. Conf. on Social Computing, 2013.
A. Abbasi, Y. Zhou, S. Deng, and P. Zhang, “Text Analytics to Support Sense-making in Social Media: A Language-Action Perspective,” MIS Quarterly, vol. 42, no. 2, pp. 427-464, 2018.
D. Adjeroh, R. Beal, A. Abbasi, W. Zheng, M. Abate, and A. Ross, "Signal Fusion for Social Media Analysis of Adverse Drug Events," IEEE Intelligent Systems, vol. 29, no. 2, pp. 74-80, 2014.
C. Adrover, T. Bodnar, and M. Salathe, "Targeting HIV-Related Medication Side Effects and Sentiment Using Twitter Data," arXiv Preprint, 2014.
A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data," Proc. of ACL HLT Conf., pp. 30-38, 2011.
AiApplied, www.ai-applied.nl, 2015.
Alexa.com, "Website Traffic Ranking," www.alexa.com, 2015.
G. Amati, M. Bianchi, and G. Marcone, "Sentiment Estimation on Twitter," IIR, pp. 39-50, 2014.
N. Aston, J. Liddle, and W. Hu, "Twitter Sentiment in Data Streams with Perceptron," Journal of Computer and Communications, 2014.
A. Bakliwal, J. Foster, J. van der Puil, R. O'Brien, L. Tounsi, and M. Hughes, "Sentiment Analysis of Political Tweets: Towards an Accurate Classifier," Proc. of ACL Workshop on Language in Social Media, pp. 49-58, 2013.
L. Barbosa and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," Proc. of Intl. Conf. on Computational Linguistics, pp. 36-44, 2010.
A. Bermingham and A. Smeaton, "Classifying Sentiment in Microblogs: Is Brevity an Advantage?," Proc. of ACM CIKM Conf., pp. 1833-1836, 2010.
S. Bhuta, A. Doshi, U. Doshi, and M. Narvekar, "A Review of Techniques for Sentiment Analysis of Twitter Data," Proc. of Intl. Conf. on Issues and Challenges in Intelligent Computing Techniques, 2014.
A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data," Proc. of Intl. Conf. on Discovery Science, pp. 1-15, 2010.
J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
F. Bravo-Marquez, M. Mendoza, B. Poblete, "Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis," Proc. of Intl. Workshop on Issues of Sentiment Discovery and Opinion Mining, 2013.
C. Callison-Burch and M. Dredze, "Creating Speech and Language Data with Amazon's Mechanical Turk," Proc. of NAACL HLT Workshop, pp. 1-12, 2010.
ChatterBox, www. chatterbox.co, 2015.
J. Chung and E. Mustafaraj, "Can Collective Sentiment Expressed on Twitter Predict Political Elections?," Proc. of AAAI Conf. on Artificial Intelligence, pp. 1770-1771, 2011.
T.S. Clark, J.K. Staton, E. Agchtein, and Y. Wang, "Revealed Public Opinion on Twitter: The Supreme Court of the United States Same-Sex Marriage Decisions," Emory University Working Paper, 2014.
A. Cui, M. Zhang, Y. Liu, and S. Ma, "Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis," Information Retrieval Technology, pp. 238-249, 2011.
S. Das and M. Chen, Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," Management Science, vol. 53, no. 9, pp. 1375-1388, 2007.
K. Dave, S. Lawrence, and D. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," Proc. of WWW Conf., pp. 519-528, 2003.
D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," Proc. of COLING Conf., pp. 241-249, 2010.
N.A. Diakopoulos and D.A. Shamma, "Characterizing Debate Performance via Aggregated Twitter Sentiment," Proc. of SIGCHI Conf. on Human Factors in Computing Systems, pp. 1195-1198, 2010.
L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive Recursive Neural Network for Target-Dependent Twitter Sentiment Classification," Proc. of ACL Conf., pp. 49-54, 2014.
A. DuVander, "Which APIs are Handling Billions of Requests Per Day?," Programmable Web, 2012.
J. Fernández, Y. Gutiérrez, J.M. Gómez, and P. Martinez-Barco, "GPLSI: Supervised Sentiment Analysis in Twitter using Skipgrams," Proc. of ACL Workshop on Semantic Evaluation, pp. 294-299, 2014.
C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the Relationships Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," Info. Systems Research, vol. 19, no. 3,

2008.
M. Gamon, "Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis," Proc. of Conf. on Computational Linguistics, 2004.
D. Gayo-Avello, "A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter Data," Social Science Computer Review, 2013.
M. Ghiassi, J. Skinner, and D. Zimbra, "Twitter Brand Sentiment Analysis: A Hybrid System using N-gram Analysis and Dynamic Artificial Neural Network," Expert Systems with Applications, vol. 40, no. 16, pp. 6266-6282, 2013.
M. Ghiassi, D. Zimbra, and S. Lee, "Targeted Twitter Sentiment Analysis for Brands using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks," Journal of Management Information Systems, vol. 33, no. 4, pp. 1034-1058, 2016.
A. Giachanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," ACM Computing Surveys, vol. 49, no. 2, 2016.
B. Gleason, "#Occupy Wall Street: Exploring Informal Learning about a Social Movement on Twitter," American Behavioral Scientist, 2013.
A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," Stanford Digital Library Technologies Project Technical Report, 2009.
P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and Combining Sentiment Analysis Methods," Proc. of ACM Conf. on Online Social Networks, pp. 27-38, 2013.
T. Gunther and L. Furrer "GU-MLT-LT: Sentiment Analysis of Short Messages using Linguistic Features and Stochastic Gradient Descent," Proc. of the Intl. Workshop on Semantic Evaluation, 328-332, 2013.
M. Hagan, M. Potthast, M. Buchner, and B. Stein "Webis: An Ensemble for Twitter Sentiment Detection," Proc. of the Ninth International Workshop on Semantic Evaluation, 582-589, 2015.
M. A. K. Halliday, An Introduction to Functional Grammar, (3rd ed.), London: Hodder Arnold, 2004.
A. Hassan, A. Abbasi, and D. Zeng, "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework," Proc. of ASE / IEEE Intl. Conf. on Social Computing, pp. 357-364, 2013.
W. Hu, "Real-Time Twitter Sentiment toward Thanksgiving and Christmas Holidays," Social Networking, vol. 2, pp. 77-86, 2013.
Y. Hu, F. Wang, and S. Kambhampati, "Listening to the Crowd: Automated Analysis of Events via Aggregated Twitter Sentiment," Proc. of Conf. on Artificial Intelligence, pp. 2640-2646, 2013.
X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised Sentiment Analysis with Emotional Signals," Proc. of WWW Conf., pp. 607-618, 2013.
S. Huang, W. Peng, J. Li, and D. Lee, "Sentiment and Topic Analysis on Social Media: A Multi-Task Multi-Label Classification Approach," Proc. of ACM WebSci Conf., 2013.
Intridea, www. intridea.com, 2015.
B. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth," J. of the Amer. Soc. for Info. Sci. and Tech., vol. 60, no. 11, pp. 2169-2188, 2009.
X. Ji, S.A. Chun, and J. Geller, "Monitoring Public Health Concerns Using Twitter Sentiment Classifications," Proc. of Healthcare Informatics Conf., pp. 335-344, 2013.
L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-Dependent Twitter Sentiment Classification," Proc. of ACL Conf., pp. 151-160, 2011.
F. Jiang, Y. Liu, H. Luan, M. Zhang, and S. Ma, "Microblog Sentiment Analysis with Emoticon Space Model," Social Media Processing, Springer, pp. 76-87, 2014.
F. Jungermann, "Information Extraction with Rapidminer," Proc. of GSCL Conf., pp. 50-61, 2009.
Y. Kaewpitakkun, K. Shirai, and M. Mohd, "Sentiment Lexicon Interpolation and Polarity Estimation of Objective and Out-Of-Vocabulary Words to Improve Sentiment Classification on Microblogging," Proc. of Pacific Asia Conf. on Language, Information, and Computation, pp. 204-213, 2014.
F.H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter Opinion Mining Framework using Hybrid Classification Scheme," Decision Support Systems, vol. 57, pp. 245-257, 2014.
S. Kim and E. Hovy, "Determining the Sentiment of Opinions," Proc. of Intl. Conf. on Computational Linguistics, pp. 1-8, 2004.
D. Kim and J.W. Kim, "Public Opinion Mining on Social Media: A Case Study of Twitter Opinion on Nuclear Power," Proc. of CES-CUBE, 2014.
E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-Based Sentiment Analysis of Twitter Posts," Expert Systems with Applications, 2013.
E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," Proc. of AAAI Conf. on Weblogs and Social Media, pp. 538-541, 2011.
Y. Liu, "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," Journal of Marketing, vol. 70, pp. 74-89, 2006.
K. Liu, W. Li, and M. Guo, "Emoticon Smoothed Language Models for Twitter Sentiment Analysis," Proc. of AAAI Conf., 2012.
S. Liu, X. Cheng, and F. Li, "TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets," IEEE Trans. on Knowledge and Data Engineering, vol. 27, no. 6, 2015.

Lymbix, www.lymbix.com, 2015.
E. Marinez-Camara, M. Martin-Valdivia, L. Urena-Lopez, and A. Montejo-Raez, "Sentiment Analysis in Twitter," Natural Language Engineering, vol. 20, no. 1, 2012.
J. R. Martin and P. R. White, The Language of Evaluation: Appraisal in English, 2005.
E. Mayfield and C.P. Rosé, "LightSIDE: Open Source Machine Learning for Text Accessible to Non-Experts," Handbook of Automated Essay Grading, 2012.
Y. Mejova, P. Srinivasan, and B. Boynton, "GOP Primary Season on Twitter: Popular Political Sentiment in Social Media," Proc. of ACM WSDM Conf., 2013.
A. Mittal and A. Goel, "Stock Prediction Using Twitter Sentiment Analysis," Stanford University Working Paper, 2012.
Y. Miura, S. Sakaki, K. Hattori, and T. Ohkuma "TeamX: A Sentiment Analyzer with Enhanced Lexicon Mapping and Weighting Scheme for Unbalanced Data," Proc. of the Eighth Intl. Workshop on Semantic Evaluation, 628-632, 2014.
MLAnalyzer, www.mashape.com/mlanalyzer, 2015.
S.M. Mohammad, S. Kiritchenko, and X. Zhu, X. "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," Proc. of the Seventh International Workshop on Semantic Evaluation, 321-327, 2013.
A. Montejo-Ráez, E. Martínez-Cámara, M.T. Martín-Valdivia, and L.A. Ureña-López, "Ranked WordNet Graph for Sentiment Polarity Classification in Twitter," Computer Speech and Language, vol. 28, no. 1, pp. 93-107, 2014.
F. Nielsen, "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs," Proc. of ESWC Workshop on Making Sense of Microposts, 2011.
B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," Proc. of AAAI Conf. on Weblogs and Social Media, pp. 122-129, 2010.
A. Ortigosa, J.M. Martin, and R.M. Carro, "Sentiment Analysis in Facebook and its Application to E-Learning," Computers in Human Behavior, vol. 31, pp. 527-541, 2014.
A. Pak and P. Paroubek, "Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives," Proc. of ACL Workshop on Semantic Evaluation, pp. 436-439, 2010a.
A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Proc. of LREC, pp. 1320-1326, 2010b.
B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification using Machine Learning Techniques," Proc. of EMNLP Conf., pp. 79-86, 2002.
T. Proisl, P. Greiner, S. Evert and B. Kabashi "Simple and Robust Methods for Polarity Classification," Proc. of the Seventh Intl. Workshop on Semantic Evaluation, 395-401, 2013.
T. Rao and S. Srivastava, "Twitter Sentiment Analysis: How to Hedge Your Bets in the Stock Markets," State of the Art Applications of Social Network Analysis, Springer, pp. 227-247, 2014.
Repustate, www.repustate.com, 2015.
E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," Proc. of EMNLP Conf., pp. 105-112, 2003.
M. Ringsquandl and D. Petkovic, "Analyzing Political Sentiment on Twitter," AAAI Spring Symposium: Analyzing Microtext, pp. 40-47, 2013.
H. Rui, Y. Liu, and A. Whinston, "Whose and What Chatter Matters? The Effect of Tweets on Movie Sales," Decision Support Systems, vol. 55, no. 4, pp. 863-870, 2013.
H. Saif, M. Fernandez, Y. He, and H. Alani, "Senticircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter," Proc. of Extended Semantic Web Conf., 2014a.
H. Saif, M. Fernandez, Y. He, and H. Alani, "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter," Proc. of LREC, 2014b.
H. Saif, Y. He, and H. Alani, "Alleviating Data Sparsity for Twitter Sentiment Analysis," Proc. of ACM Intl. WWW Conf., pp. 2-9, 2012.
N. Sanders, "Twitter Sentiment Corpus," Sanders Analytics 2.0, 2011.
Scherer, K. R. Appraisal Theory, New York, NY, US: John Wiley & Sons Ltd, 1999.
Scherer, K. R. "What are Emotions? And How Can They Be Measured?" Social Science Information, vol. 44, no. 4, pp. 695-729, 2005.
Semantria, www.semantria.com, 2015.
SentimentAnalyzer, www.sentimentanalyzer.appspot.com, 2015.
H. Sharif, A. Abbasi, F. Zaffar, and D. Zimbra, "Detecting Adverse Drug Reactions using a Sentiment Classification Framework," Proc. of ASE / IEEE Intl. Conf. on Social Computing, 2014.
A. Siganos, E. Vagenas-Nanos, P. Verwijmeren, "Facebook's Daily Sentiment and International Stock Markets," Journal of Economic Behavior & Organization, vol. 107, pp. 730-743, 2014.
J. Smailovic, M. Grcar, N. Lavrac, and M. Žnidaršic, "Predictive Sentiment Analysis of Tweets: A Stock Market Application," Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Springer, pp. 77-88, 2013.

R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank," Proc. of EMNLP Conf., 2013.
M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph," Proc. of EMNLP Conf., pp. 53-63, 2011.
S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, "Interpreting the Public Sentiment Variations on Twitter," IEEE Trans. on Knowledge and Data Engineering, vol. 6, no. 1, 2012.
D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," Proc. of ACL Conf., pp. 1555-1565, 2014.
P.A. Tapia and J.D. Velásquez, "Twitter Sentiment Polarity Analysis: A Novel Approach for Improving the Automated Labeling in a Text Corpora," Active Media Technology, pp. 274-285, 2014.
P. Tetlock, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," The Journal of Finance, vol. 62, pp. 1139-1168, 2007.
TextProcessing, www.text-processing.com, 2015.
Textalytics, www.textalytics.com, 2015.
M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment Strength Detection in Short Informal Text," Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544-2558, 2010.
M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in Twitter Events," Journal of the American Society for Information Science and Technology, vol. 62, no. 2, pp. 406-418, 2011.
C. Troussas, M. Virvou, K.J. Espinosa, K. Llaguno, and J. Caro, "Sentiment Analysis of Facebook Statuses using Naïve Bayes Classifier for Language Learning," Proc. Intl. Conf. in Information, Intelligence, Systems and Applications (IISA), pp. 1-6, 2013.
A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," Proc. of AAAI Conf. on Weblogs and Social Media, pp. 178-185, 2010.
P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proc. of Annual Meeting of ACL, pp. 417–424, 2002.
Twitter, Inc., "IPO Prospectus," http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm, 2014.
Twitter, Inc., "Second Quarter 2016 Report," https://investor.twitterinc.com/results.cfm, 2016.
UClassify, www.uclassify.com, 2015.
A. Vanzo, D. Croce, and R. Basili, "A Context-Based Model for Sentiment Analysis in Twitter," Proc. of COLING Conf., pp. 2345-2354, 2014.
A. Verma, K.A.P. Singh, and K. Kanjilal, "Knowledge Discovery and Twitter Sentiment Analysis: Mining Public Opinion and Studying its Correlation with Popularity of Indian Movies," International Journal of Management, vol. 6, no. 1, pp. 697-705, 2015.
ViralHeat, www.viralheat.com, 2015.
H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle," Proc. of ACL Conf., 2012.
C. Whitelaw, N. Garg, and S. Argamon, "Using Appraisal Groups for Sentiment Analysis," In Proc. of 14th ACM International Conference on Information and Knowledge Management, pp. 625-631, 2005.
J. Wiebe, T. Wilson, and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," Language Resources and Evaluation, vol. 39, no. 2-3, pp. 165-210, 2005.
T. Winograd, "A Language/Action Perspective on the Design of Cooperative Work," In Proc. of ACM Conference on Computer-supported Cooperative Work, pp. 203-220, 1986.
I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2005.
B. Xiang and L. Zhou, "Improving Twitter Sentiment Analysis with Topic-Based Mixture Modeling and Semi-Supervised Training," Proc. of ACL Conf., pp. 434-439, 2014.
W. Zhang and S. Skiena, "Trading Strategies to Exploit Blog and News Sentiment," Proc. of ICWSM, 2010.
L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis," Hewlett-Packard Labs Technical Report, 2011.
Z. Zhou, X. Zhang, and M. Sanderson, "Sentiment Analysis on Twitter through Topic-Based Lexicon Expansion," Databases Theory and Applications, pp. 98-109, 2014.
D. Zimbra, H. Chen, and R.F. Lusch, "Stakeholder Analyses of Firm-Related Web Forums: Applications in Stock Return Prediction," ACM Trans. on Management Information Systems, vol. 6, no. 1, 2015.