Benchmarking Twitter Sentiment Analysis Tools

Conference Paper · May 2014

CITATIONS
59

READS
7,055

3 authors:

Ahmed Abbasi
University of Notre Dame
75 PUBLICATIONS **3,455** CITATIONS
SEE PROFILE

Ammar Hassan
University of Virginia
2 PUBLICATIONS **166** CITATIONS
SEE PROFILE

Milan Dhar
University of Virginia
1 PUBLICATION **59** CITATIONS
SEE PROFILE

Benchmarking Twitter Sentiment Analysis Tools

Ahmed Abbasi, Ammar Hassan, Milan Dhar
University of Virginia
Charlottesville, Virginia, USA
E-mail: abbasi@comm.virginia.edu, mah9tg@virginia.edu, msd4ah@virginia.edu

Abstract

Twitter has become one of the quintessential social media platforms for user-generated content. Researchers and industry practitioners are increasingly interested in Twitter sentiments. Consequently, an array of commercial and freely available Twitter sentiment analysis tools have emerged, though it remains unclear how well these tools really work. This study presents the findings of a detailed benchmark analysis of Twitter sentiment analysis tools, incorporating 20 tools applied to 5 different test beds. In addition to presenting detailed performance evaluation results, a thorough error analysis is used to highlight the most prevalent challenges facing Twitter sentiment analysis tools. The results have important implications for various stakeholder groups, including social media analytics researchers, NLP developers, and industry managers and practitioners using social media sentiments as input for decision-making.

Keywords: sentiment analysis, twitter, social media analytics, opinion mining, benchmarking

1. Introduction

Twitter has emerged as one of the premier social media analytics channels. With over 3 billion tweets and 15 billion API calls generated daily (DuVander, 2012), Twitter has an abundance of both supply and demand. The demand is partially precipitated by the growing body of social media analytics applications involving Twitter. One huge area is Twitter sentiments, which are used for understanding consumer perceptions (Smith et al., 2012), predicting financial performance (Bollen et al., 2011), providing early warnings for adverse medical events (Abbasi et al., 2013; Fu et al., 2012), determining and understanding election outcomes (Skoric et al. 2011), and as input for disaster response surveillance systems (Goodchild & Glennon, 2010). In all of these examples, and many others like them, the key sentiment input is whether a given tweet has positive, negative, or neutral sentiment polarity regarding the target of interest (Abbasi et al., 2008a). These inputs are often aggregated to develop social media sentiment signals or indexes over time, with time series ranging from strong positive to strong negative. The quality of tweet sentiment polarity classifications is hence a critical intermediary step in the information value chain and a big part of emerging real-time analytics applications. Accordingly, numerous commercial and freely available tools have emerged for Twitter sentiment classification. One study estimated the number at over 50 tools, with new commercial start-ups and academic offerings arising on a monthly basis. What remains unclear is how effective the plethora of available tools really is. Extensive benchmarking is warranted for a couple of reasons:

- In order to provide potential users of Twitter sentiment analysis techniques with a "consumer report" of existing tools.
- In order to assess the state-of-the-art for Twitter sentiment analysis and examine possible avenues for improvement.

Accordingly, with these objectives in mind, we performed a large-scale benchmark study of Twitter sentiment analysis tools. We examined the sentiment polarity classification performance of 20 tools, including both commercial and freely available offerings, on 5 carefully crafted Twitter test beds annotated using Amazon Mechanical Turk. The results revealed that tool performances varied considerably, with the best-performing tools attaining overall accuracies of between 65% and 71% on average, while many low-performing tools yielded accuracies below 50%. Furthermore, tool results also varied across test beds, suggesting the presence of a domain interaction on tool performance. In order to further investigate tool performances, detailed error analysis was conducted on the most frequently misclassified tweets across the five test beds. We manually examined the most prevalent misclassifications and developed a two-level hierarchical error taxonomy. The taxonomy, and prevalence of certain error types within the taxonomy, were both used to make inferences about the strengths and weaknesses of the 20 tools examined. The results have important implications for several stakeholder groups, including social media analytics researchers, natural language processing (NLP) and text mining researchers and developers, and industry practitioners that utilize Twitter sentiment signals as input for decision-making.

2. Twitter Sentiment Analysis Tools

Many tools have been developed in recent years for analyzing sentiments in short informal social media texts. We examined a fairly diverse set of 20 Twitter sentiment analysis tools. These tools included freely available systems developed in academic settings, commercial API-based tools requiring a monthly subscription, and even a few algorithms published in the NLP literature. While the 20 included are by no means an exhaustive list, they include tools with hundreds of paying customers, ones that have been downloaded several thousand times, ones appearing in papers cited hundreds of times, and

even ones from large companies with millions of daily API calls. Details regarding the tools are as follows.

The 20 tools evaluated can be broadly grouped into two categories: stand-alone commercial tools and trained workbench tools. The stand-alone tools use text analytics models that can be applied directly to unlabeled documents immediately "out-of-the-box." These tools include API-based offerings and ones that can be downloaded as desktop applications. We incorporated 15 such tools in the evaluation. The ready-to-use nature of these tools, without the need for domain-specific model development or training, makes them easier to apply "off-the-shelf." However, the lack of domain-specificity can also be detrimental from a performance perspective since the tools' underlying models may incorporate rules/assumptions that are erroneous or inapplicable in the context of a particular test bed. The stand-alone commercial tools evaluated were uClassify, ChatterBox, Sentiment140 (Go et al. 2009), Textalytics, Intridea, AiApplied, ViralHeat, Lymbix, Anonymous (the terms of use for this tool prevented us from mentioning it by name), SentimentAnalyzer, TextProcessing, Semantria, SentiStrength (Thelwall et al., 2010), MLAnalyzer, and Repustate. Most of the stand-alone tools incorporated in the study are commercial offerings accessed either directly through the vendor's API or through a third-party API marketplace such as MashApe. Two exceptions are Sentiment140 and SentiStrength, both of which were developed as a result of published academic research.

SentiStrength (Thelwall et al., 2010) is a popular stand-alone sentiment analysis tool. It uses a sentiment lexicon for assigning scores to negative and positive phrases in text. To determine sentence or document level polarities, the phrase level scores can be aggregated. The unsupervised-learning nature of such an approach makes it easily applicable to any data set. Sentiment140 uses a trained machine learning classifier built on a large Twitter corpora of positive and negative tweets automatically developed based on the presence of emoticons (Go et al. 2009). The tool uses word and part-of-speech tag n-grams coupled with a maximum entropy-based machine learning classifier.

Some of the stand-alone tools only output continuous polarity scores as opposed to discrete polarity classifications. For such tools (e.g., ChatterBox), we discretized the continuous scores into three categories (for positive, neutral, and negative) using bins that maximized the tools' performance with respect to overall accuracy and class-level recalls. While such an approach might have inflated the performance of certain stand-alone tools, we felt that maximizing the performance of tools to the extent possible embodied a user-centric perspective.

The workbench tools are ones that require supervised learning-based model development on a labeled training set. These provide options for stemming, tokenizing, inclusion of different feature representations, and parameters for number of features to incorporate in the models. The 5 workbench tools included were LightSide, BPEF, EWGA, FRN, and a word n-gram baseline run using the text processing extension in RapidMiner (Abbasi et al., 2008; Mayfield & Rose, 2012; Hassan et al., 2013; Kouloumpis et al., 2011; Abbasi et al., 2011; Jungermann 2009). Workbench tools require extensive parameter tuning and validation in a training environment, but have the potential to incorporate task and domain-specific knowledge.

EWGA uses an entropy-weighted genetic algorithm to efficiently select features for sentiment classification using a wrapper-model, where the performance of a feature subset is used as its fitness function value within the genetic algorithm (Abbasi et al., 2008a). FRN uses a feature relation network comprising of two key syntactic n-gram relations: subsumption and parallel relations (Abbasi, 2010; Abbasi et al. 2011). These relations are used to efficiently perform feature selection from rich feature spaces encompassing many different types of n-grams. The reduced feature set is input into an SVM classifier.

BPEF utilizes a bootstrap parametric ensemble framework (Hassan et al., 2013). An ensemble encompassing tens of thousands of binary one-against-one classifiers are constructed leveraging different combinations of data sets, feature set combinations, and machine learning classifiers (e.g., Information, Bayesian, and/or Statistical Learning Theory-based). A meta-level search heuristic is used to identify a small subset of models ultimately retained for classification. The word n-gram baseline comprised of unigrams, bigrams, and trigrams selected using the information gain heuristic and coupled with an SVM classifier, as done in prior studies (Pang and Lee, 2002; Abbasi et al., 2008b; Abbasi & Chen, 2008).

3. Evaluation Test Bed and Metrics

We included tweets pertaining to 5 broad topics: telecommunications, pharmaceutical, security, technology, and retail consumer goods. The topics chosen are relevant to an array of application areas, including consumer sentiments, social media for smart health, security informatics, and user experiences. All data sets were labeled with gold standard sentiment polarities using Amazon Mechanical Turk (AMT). The technology data set was developed by Sanders (2011). For the remaining four, we followed best practices for data annotation previously outlined (Callison-Burch & Dredze, 2010). Prior to using AMT, manual and automated pre-processing methods were used to remove irrelevant tweets (e.g., non-English, or unrelated to the topics of interest). Within AMT, the Sentiment Rating module was employed, using 5 experienced turks per tweet. Furthermore, only tweets with sentiments relevant to the targets were labeled as positive or negative.

Data Set	Description of Tweets	Quantity of Tweets			
		Overall	Positive	Negative	Neutral
Telco	Related to telecommunications company Telus' products, and services. Include general discussion of experiences, news, and specific events.	5281	20.9%	8.9%	70.2%
Pharma	Related to users' experiences with pharmaceutical drugs. Include mentions of adverse events, positive interactions, etc.	5009	15.6%	11.1%	73.3%
Security	Related to major security companies' products and services, including security incidents and new software releases and/or security patches.	5086	24%	11.1%	64.9%
Tech	Related to four major tech firms. Include discussion of companies' products, services, policies, and general user experiences.	3502	15.1%	16.9%	68.0%
Retail	Include discussion of a specific category of retail products (household paint) and user experiences related to those products.	3750	42.7%	9.0%	48.3%

Table 1: Evaluation Test Bed Overview

Table 1 provides an overview of the test bed. Not surprisingly, the majority of tweets in each test bed were neutral, since most social media communications associated with a given topic do not contain positive or negative sentiment. The one exception was the Retail test bed, where people discussing household paints tended to have predominantly either positive or negative sentiments. Nevertheless, the five data sets were all quite imbalanced with respect to distributions of tweets across positive, negative, and neutral polarity classes.

Several standard evaluation metrics were employed. These included overall accuracy and class-level recall and class-level precision. Overall accuracy is the percentage of total tweets classified correctly (as positive, neutral, or negative). Class-level recall is the percentage of tweets associated with a particular class that were classified as such. For example, negative recall is the percentage of all negative tweets in the test bed that are classified as negative. Class-level precision is the percentage of all tweets classified as belonging to a given class that actually belong to that class. In the paper, due to space constraints, only the results for overall accuracy and class-level recall rates are reported.

4. Experiment Results

Table 2 presents the experiment results for the 15 stand-alone commercial tools, while Table 3 depicts overall accuracies for the 5 workbench methods. Also included in the tables is the average accuracy across the 5 test beds, as an indicator of overall performance. With respect to stand-alone tools, SentiStrength, ChatterBox, Sentiment140, and Texalytics provided the best overall performance across test beds, with average accuracies above 66%. Amongst the top-performing stand-alone tools, Sentiment140's performance was the most balanced across test beds, with accuracies ranging from 61% to 71%. On the other hand, four of the tools' average accuracies were below 50%, and median average accuracy across the 15 tools was only 56%.

These results suggest that the quality of Twitter sentiment analysis tools varies considerably, and that this performance variation can have important implications for various Twitter-based social media analytics applications. Based on the results presented in Table 3, not surprisingly, the workbench methods provided higher average accuracy across test beds since they were trained on similar data as the evaluation sets. These methods' average accuracies were between 67% and 71%, and with generally less variation across test beds (e.g., BPEF accuracies varied only 6% between test beds). The results in Tables 2 and 3 illustrate the possible trade-offs between using stand-alone and workbench tools.

In order to further investigate these potential trade-offs, we examined the class-level recall rates for the top-performing stand-alone and workbench tools. The top five stand-alone and top three workbench tools were incorporated. These results are presented in Tables 4 and 5. From the tables, it is evident that class-level recall rates varied considerably for certain tools, across data sets. With respect to stand-alone tools, SentiStrength attained good overall accuracies due to higher positive and negative recall rates on data sets such as Pharma, Security, and Telco (ranging between 80% and 90% on negative recall). However, these higher positive/negative rates were accompanied by markedly lower neutral-class recall rates (i.e., many tweets with falsely positive/negative sentiment attributions). Similarly, ChatterBox tended to attain higher recall on positive and neutral tweets, whereas Textalytics was most effective in terms of neutral recall (between 86% and 90% across test beds).

With respect to workbench tools, LightSide and FRN exhibited a similar bias towards a specific polarity class. Both methods attained higher neutral recall rates but with markedly lower positive and negative recall values. Conversely, BPEF attained relatively balanced class-level recall values across the positive, negative, and neutral classes. Its class-level recall rates were generally within

Tool	Average	Pharma	Retail	Security	Tech	Telco
SentiStrength	**67.49**	74.68	56.35	65.51	69.61	71.31
Chatterbox	67.43	**75.04**	53.19	67.20	69.73	**71.99**
Sentiment140	66.46	62.09	**61.77**	**68.84**	67.82	71.79
Textalytics	66.22	70.33	55.14	66.33	68.29	71.02
Intridea	63.31	64.18	47.37	62.63	**75.19**	67.20
AiApplied	61.84	69.59	47.99	64.05	60.39	67.20
ViralHeat	61.16	63.77	48.42	61.94	64.12	67.56
Lymbix	56.63	52.03	54.81	47.60	63.45	65.25
SentimentAnalyzer	55.15	55.33	51.36	54.83	56.50	57.75
TextProcessing	54.06	49.68	50.01	58.40	52.40	59.79
Semantria	53.50	44.68	56.33	45.46	60.99	60.06
uClassify	47.22	51.70	42.12	47.51	50.31	44.47
MLAnalyzer	45.20	37.95	52.15	41.35	48.06	46.47
Repustate	43.98	35.80	41.06	31.93	40.90	70.20
Anonymous	40.86	33.65	49.93	32.71	43.11	44.89

Table 2: Overall Accuracies for 15 Stand-Alone Tools on 5 Test Beds

Tool	Average	Pharma	Retail	Security	Tech	Telco
BPEF	**71.38**	67.81	**65.24**	**75.32**	76.30	72.21
Lightside	69.35	70.71	58.22	69.86	**76.99**	70.99
FRN	69.17	**72.60**	59.96	69.98	71.00	**72.30**
EWGA	68.12	70.21	60.00	68.50	70.50	71.41
RapidMiner	66.86	67.50	59.52	66.02	70.02	71.22

Table 3: Overall Accuracies for 5 Workbench Tools on 5 Test Beds

Tool		Pharma			Retail			Security		
		Pos.	Neg.	Neu.	Pos.	Neg.	Neu.	Pos.	Neg.	Neu.
Stand-Alone	SentiStrength	46.98	90.47	29.29	53.28	38.10	62.44	87.88	90.93	0.15
	Chatterbox	37.23	56.65	62.47	57.55	11.01	52.79	63.39	30.07	66.52
	Sentiment140	44.03	62.59	65.84	43.76	11.01	87.09	61.02	25.98	79.02
	Textalytics	28.75	23.74	86.20	21.04	10.39	93.60	24.80	16.01	90.25
	Intridea	26.19	68.71	37.45	32.52	63.69	37.34	32.92	62.23	39.89
Work bench	BPEF	63.18	61.33	69.76	60.74	64.58	69.33	75.92	72.42	75.60
	Lightside	44.93	28.06	82.63	57.24	34.23	63.54	56.59	47.33	78.60
	FRN	39.15	21.04	84.86	58.99	24.70	67.35	55.20	35.94	81.23

Table 4: Class-level Recalls for Select Stand-Alone and Workbench Tools on Pharma, Retail, and Security Test Beds

Tool		Tech			Telco		
		Pos.	Neg.	Neu.	Pos.	Neg.	Neu.
Stand-Alone	SentiStrength	62.00	45.27	61.89	59.67	80.56	42.47
	Chatterbox	52.64	45.10	56.51	53.16	33.97	64.39
	Sentiment140	55.09	36.82	78.36	46.38	28.42	84.84
	Textalytics	26.98	18.04	90.00	25.07	29.06	90.02
	Intridea	69.81	81.76	74.75	32.73	68.38	46.32
Work bench	BPEF	69.25	78.55	77.31	57.01	74.52	76.47
	Lightside	47.92	55.91	84.29	38.70	49.79	83.30
	FRN	37.17	42.06	80.92	33.30	37.26	81.10

Table 5: Class-level Recalls for Select Stand-Alone and Workbench Tools on Tech and Telco Test Beds

10% of one another, with the exception of the Telco data set, where the positive recall rate was about 17% to 19% lower than the other two classes. Class-level recall rates in general, and balance in particular, both have important implications for social media analytics applications involving sentiment. For instance, Hassan et al. (2013) showed that tools with lower positive and negative recall rates are susceptible to generating sentiment index time series that are "flatter" and less effective for representing events with extreme positive or negative sentiments. They illustrated this point with a sentiment time series for a major North American telecommunications provider.

5. Error Analysis

The performance variation between stand-alone tools' accuracies (over 26% on average), and the generally low performance of tools (mostly below 70%), both underscore the challenges associated with effective Twitter sentiment classification. In order to better understand these challenges, detailed error analysis was performed on the 5 test beds. We examined the 1000 most misclassified tweets in each test bed. Following best practices outlined in prior studies (Wiebe et al. 2005), for each tweet, multiple annotators categorized the errors. More specifically, three annotators classified the erroneous tweets. After multiple rounds of discussion and annotation, the errors were grouped into a two-level hierarchical taxonomy (presented in Figure 1). The taxonomy contains 13 top-level categories.

The misinterpreted user purpose category includes neutral-sentiment questions or requests that are mistaken for complements or criticisms. The semantics/sentence structure category includes jokes, sarcasm, rhetoric, and related literary devices that have been well-documented as being problematic for sentiment analysis tools. Misinterpreted user-purpose includes requests or questions (e.g., "It would be great if we could…"). Parsing issues include sentiments expressed in the hash tags, which are often not properly parsed. The target marketing tactics category includes events, contests, and advertisements, which are generally considered neutral by human annotators. The exception to usual sentiment cues category includes errors attributable to the presence of terms that are used in a connotation that is atypical, such as the use of curse words to indicate a positive outcome.

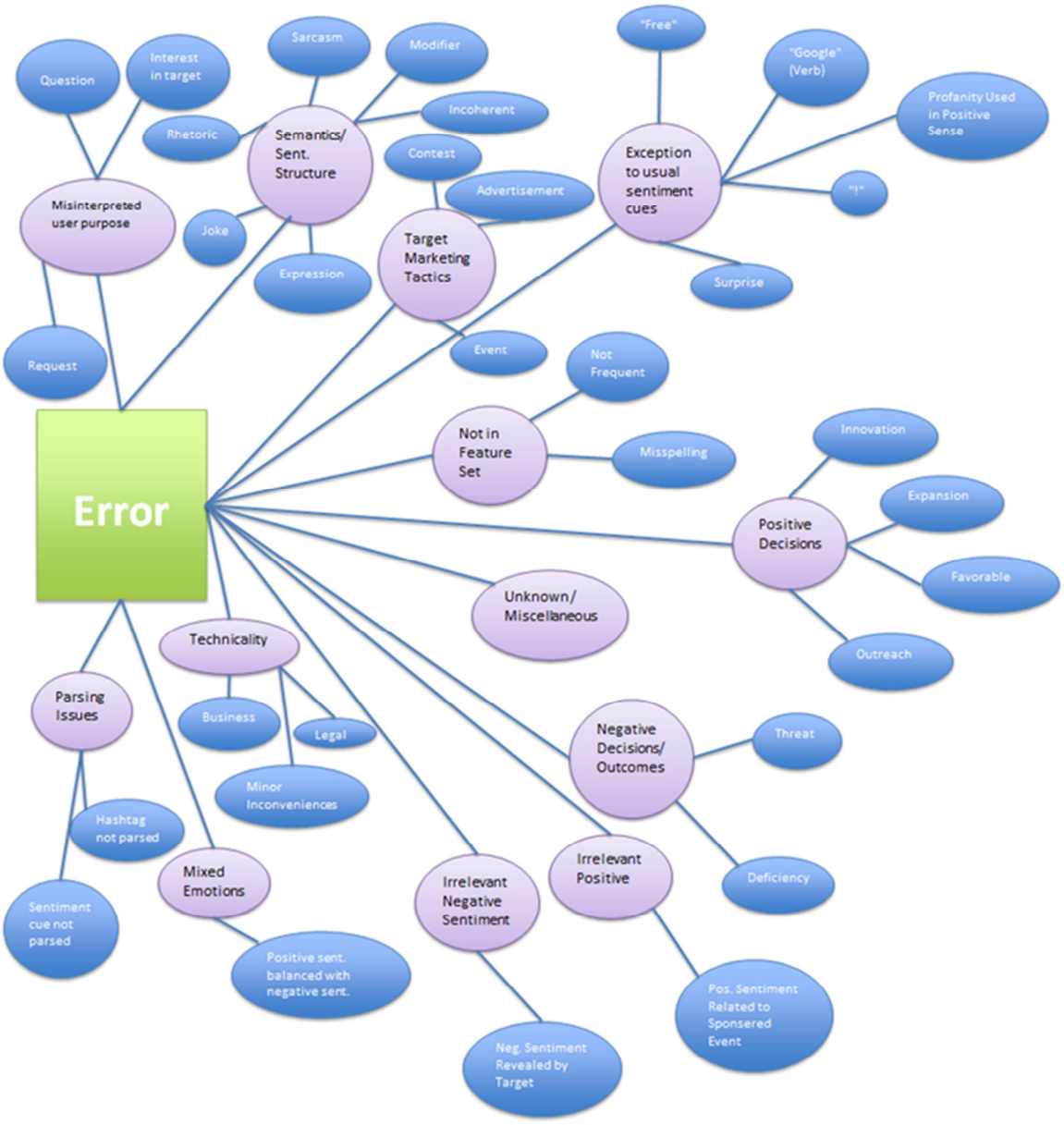

Figure 1: Twitter Sentiment Error Analysis Taxonomy

Other common categories of errors included tweets containing mixed sentiments, where the authors incorporate both positive and negative sentiments about different topics within the 140 characters. Similarly, lack of relevance of the sentiment expressed to the targets was another cause of both positive and negative misclassifications. The positive decisions category of errors was interesting. These were tweets containing subtle positive sentiment cues such as mentions of donations, charities, and other events or activities with a somewhat implied positive connotation. The tools failing to identify such tweets were generally those that presumably lacked a lexicon of positive action terms and/or keywords.

Using this taxonomy, we examined the error frequencies for the 20 tools by category, across the 5 test beds. The results are presented in Figure 2. It is apparent that certain types of errors were most prevalent. Not surprisingly, semantics/sentence structure issues (e.g., sarcasm, modifiers, jokes, rhetoric, etc.) accounted for the largest percentage of highly erroneous tweets for many test beds. This category consistently encompassed 10%-15% of total errors. However, several other categories were also quite pervasive. The positive decisions by target category accounted for over 10% of errors on four out of five data sets. The irrelevant positive and negative sentiment categories (with respect to the targets) were also responsible for between 5% and 15% of errors, each.

Interestingly, errors pertaining to mixed sentiments were very rare, despite constituting a major problem in other social media channels such as web forums and blogs. This finding suggests that the 140-character limit presents some limitations on users' abilities to articulate complex opinions encompassing multiple opposing sentiments.

Furthermore, there was an interaction between error category and data set, with certain errors being common within select domains. In most situations, the top 2-3 error categories accounted for the majority of errors on that particular data set. For example, misinterpreted user purpose was a significant source of errors on the Pharma and Retail data sets (over 20% and 40%, respectively), where questions about experiences with prescription drugs or quality of paints were misclassified as negative sentiments by several tools. Similarly, many tweets pertaining to charitable activities, fund-raising events, and donation drives in the Telco data set were misclassified by several tools, accounting for over 35% of errors. Even the semantics/sentence structure category, although pervasive across all five data sets, was more prevalent in the Tech and Security data sets. In these two domains, literary devices such as jokes, sarcasm, and rhetoric were far more prevalent (e.g., when referring to security software, smart phone manufacturers, or major technology firms). Overall the results shed light on how sentiment analysis tool errors are manifested across tweets pertaining to different topics.

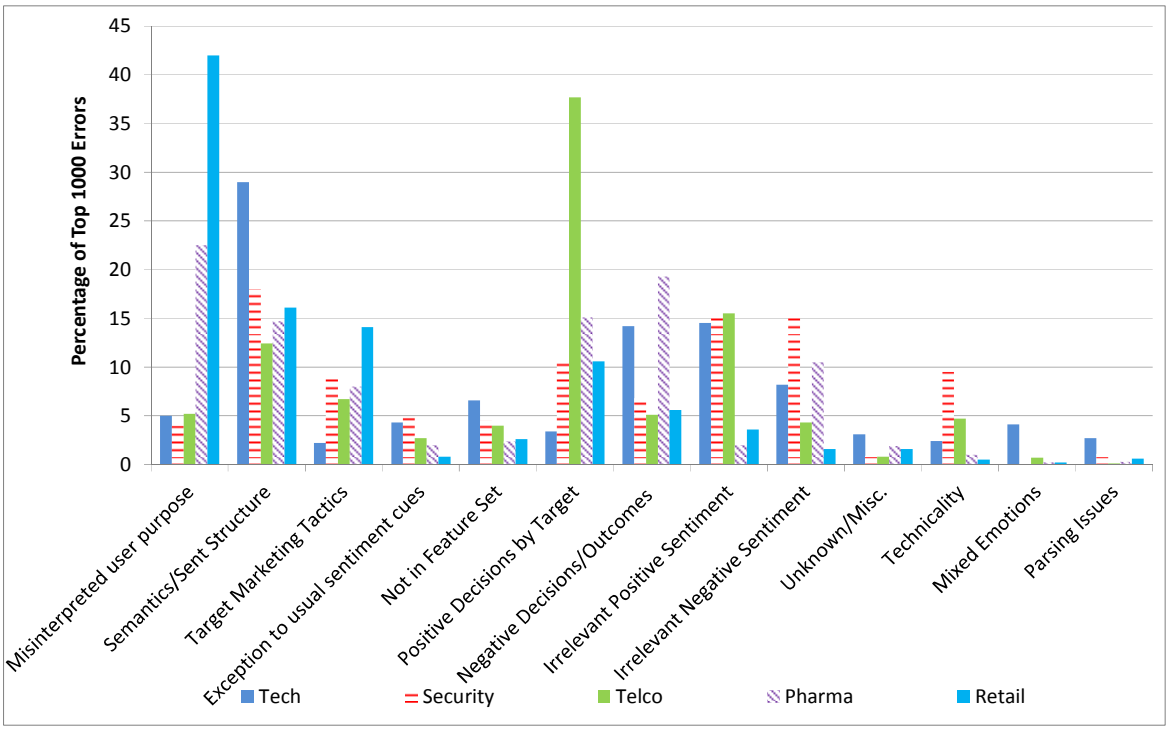

Figure 2: Prevalence of Error Types across Data Sets

6. Conclusion

The results of our work have important implications for several stakeholder groups. Social media analytics researchers can use the findings to make more informed decisions regarding their choices of specific tools for a project. They can also weigh the trade-offs between using stand-alone and workbench tools. NLP and text mining researchers and developers can use the error analysis results to improve future commercial stand-alone tools. Industry managers can be better aware of the possible strengths and weaknesses of the underlying text analytics, and how it might impact the quality and reliability of social media inputs used for decision-making. As an additional resource, some of the labeled test beds with error analysis annotations have been made publicly available through LRE Map. This will facilitate future benchmarking. Additionally, the error analysis annotation data can help guide future Twitter sentiment analysis algorithm development.

7. Acknowledgements

We would like to thank the U.S. National Science Foundation for their support of this work through the following grant: IIS-1236970.

8. References

Abbasi, A., and Chen, H. (2008). CyberGate: A Design Framework and System for Text Analysis of Computer Mediated Communication. *MIS Quarterly*, 32(4), pp. 811--837.

Abbasi, A. (2010). Intelligent Feature Selection for Opinion Classification. *IEEE Intelligent Systems*, 25(4), pp. 75--79.

Abbasi, A., Chen, H., and Salem, A. (2008a). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems*, 26(3), no. 12.

Abbasi, A., Chen, H., Thoms, S., and Fu, T. (2008b). Affect Analysis of Web Forums and Blogs using Correlation Ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 20(9), pp. 1168--1180.

Abbasi, A., France, S., Zhang, Z., and Chen, H. (2011). Selecting attributes for sentiment classification using feature relation networks. *IEEE Transactions on Knowledge and Data Engineering*, 23(3), pp. 447--462.

Abbasi, A., Fu, T., Zeng, D., and Adjeroh, D. (2013). Crawling Credible Online Medical Sentiments for Social Intelligence. *Proceedings of the ASE/IEEE International Conference on Social Computing*.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp. 1--8.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk, *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 1--12.

DuVander, A. (2012). Which APIs are Handling Billions of Requests Per Day? *Programmable Web*, May.

Fu, T., Abbasi, A., Zeng, D., and Chen, H. (2012). Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers. *ACM Transactions on Information Systems*, 30(4), no. 24.

Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224 Project Report*, Stanford, pp. 1--12.

Goodchild, M. F. and Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3), pp. 231--241.

Hassan, A., Abbasi, A., and Zeng, D. (2013). Twitter Sentiment Analysis: A Bootstrap Ensemble Framework, *Proceedings of the ASE/IEEE International Conference on Social Computing*, pp. 357--364.

Jungermann, F. (2009). Information extraction with rapidminer. *In Proceedings of the GSCL Symposium Sprachtechnologie und eHumanities*, pp. 50--61.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Mayfield, E. and Rosé, C. P. (2012). LightSIDE: Open Source Machine Learning for Text Accessible to Non-Experts. *Invited chapter in the Handbook of Automated Essay Grading*.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL Conference on Empirical methods in natural language processing*, 10, pp. 79--86.

Sanders, N. (2011). Twitter Sentiment Corpus. *Sanders Analytics 2.0*, p. 24.

Skoric, M., Poor, N., Achananuparp, P., Lim, E. P., and Jiang, J. (2012). Tweets and votes: A study of the 2011 Singapore general election. *Proceedings of the 45th Hawaii International Conference on System Science (HICSS)*, pp. 2583--2591.

Smith, A. N., Fischer, E., and Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? *Journal of Interactive Marketing* 26(2), pp. 102--113.

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

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3), pp. 165—210.