6-2017

Separating News from Opinion in Social Media using Machine learning

Swayambhu Chatterjee
Dakota State University, spchatterjee@pluto.dsu.edu

Separating News from Opinion in Social Media using Machine learning

Swayambhu Chatterjee
Dakota State University
spchatterjee@pluto.dsu.edu

ABSTRACT

Recent growth in social media platforms has produced huge amount of user-generated content. This huge commingled data set is a collection of both “factual information” like breaking-news posted by a user, as well as individual “user’s opinion”. For information extraction applications like news media, factual data are more relevant rather than user opinions and conversely for many opinion-mining applications, user’s perceptions are more important than “factual information”. Hence, separating news from opinion offers distinct advantage for both information extraction systems and opinion mining applications. In this study, we have identified limitation of existing lexicon based, subjectivity detection techniques in noisy microblog domain. We present a novel algorithm that automatically separate news from opinions from the mixed social media corpus. We used “Twitter” as our data corpus and evaluated our algorithm with eight machine-learning classifiers against the baseline. Our algorithm demonstrates significant improvement in classification accuracy over the existing baseline.

Keywords

Data mining, machine learning, algorithm, classification, Twitter

INTRODUCTION

In the current landscape of social media blogosphere, especially in microblogs like Twitter, posts related to any events are a mixed corpus of facts and opinion. Many opinion-mining applications, irrespective of their objectives like information extraction, or polarity summarization system fetch their content from the same mixed corpus and perform various classification tasks. An application that focuses on extracting facts tends to ignore individual non-factual opinions and conversely, opinion mining services that aim to analyze users” perception about a specific entity or event, find factual disclosures less relevant for their objective. For example, “factual” data consumers like news media, policy makers, research institutes or other individuals, usually interested in knowing any newly reported case of “H1N1 influenza” epidemic, exact score of any sporting event, or battery-life of the newly launched smartphone. People's perception, sentiment, or subjective opinions are not so important for them. According to Pew Research Center’s 2016 study, nearly 62% of adult population in USA get their news from social media. Identifying and verifying new information quickly are critical for journalists who use social media. However, for many corporations or businesses, it is about assessing people’s perception towards a certain events or entities. The reported case of any epidemic has huge impact on many economic decisions, primarily for industries, that related to with hospitality and tourism industry. For example, they can decide on continuing or suspending operations from West Africa. For another example, economic crash is an observed fact, but for many retail businesses, the objective is to uncover people’s opinion towards it. Decisions regarding expansion or contraction of product or services in certain demographics or market segment depend on this. In the context of comingled social media data, we aim to improve the accuracy of fact-seeking information retrieval systems and opinion mining applications by separating their respective initial feed from social media. The contributions of this work include: (1) the development of algorithm to classify factual content or news and opinion from the co-mingled social media corpus, leveraging contextual metadata and (2) demonstration of the efficacy of the algorithm by comparing it with existing lexicon-based subjectivity detection method. Experimental results indicate that our approach is highly effective in classifying news from opinions in noisy microblog domain.

LITERATURE REVIEW

Earlier work (Riloff, Wiebe and Wilson 2003; Wiebe 2002; Yu and Hatzivassiloglou 2003) has addressed this need of Opinion Identification from subjective test passages from news articles and extended reviews. However, in the context of noisy microblogs, we face certain unique challenges in dealing with short texts. These distinct features includes data sparseness (not enough words), large scale (data labeling bottlenecks), immediacy (real time - velocity), non-standardization (noisy data- misspellings) (Song, Ye, Du, Huang and Bie 2014). Traditional classification approaches using BOW (Bag of Words), TF-IDF (Term Frequency- Inverse Document Frequency) and different variation of n-grams may not perform as good as, compared to content which are formally expressed in long text format (Sriram, Fuhry, Demir, Ferhatosmanoglu and Demirbas 2010).

Social media has emerged to be news platform for many as it reaches out to a huge audience way faster than traditional media to report an event. Consequently, study related to identifying and monitoring events from social media has gained significant attention. (Atefeh and Khreich, 2013) has provided a detail literature survey on this. However, the captured events or content from social media has issues with credibility of user, description of irrelevant private events or exaggerated narrative. The objective of our study is to isolate news from opinion from social media postings. Assessing "truthfulness" of the news is beyond the scope of this study. Assuming "news" are factual information; we adopt the definition of our construct from extant literature. "Facts" are objective expressions about entities, events and their properties and "opinions" are usually subjective expressions that describe people's sentiments, appraisals or feelings toward entities, events and their properties (Liu 2010). For this classification task, the technique of "subjectivity detection" appeared to be a common approach in previous studies. It analyzes the linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs and speculations known as private states (Wiebe and Riloff 2005). Existing lexicon based subjectivity detection methods tokenize each word and compare it to a predefined set of opinionated words. However, noisy words in twitter like "grt" or "scarrry" does not have any match in that dictionary and hence often ignored. Also, though news or "factual-information" are often objective in nature, factual statements sometimes bear strong subjective cues like "horrible accident happened..". We have identified the following research gaps. At a conceptual level, we have seen that subjectivity detection technique often depends on the theme of the topic. Existing approach, identifies many news or factual-information as "subjective opinions", simply because they contain words that bear subjective cues. In addition, from an operational perspective, we have seen that socio-linguistic or lexicon-based approach may not be suitable in the context of microblogs, due to its inherent noisy nature.

RESEARCH DESIGN

Our proposed artifact is development of a novel algorithm that automatically classifies factual-information from personal opinion in microblog. We have selected Twitter as the context of our study and treated each tweet post as a single document. Along with the tweet message, we also leveraged other relevant contextual metadata of the user, like "follower - followee ratio", "user description". As part of design evaluation, we have evaluated our algorithm against leading opinion detection tool, OpinionFinder (Riloff et al. 2003; Wiebe and Riloff 2005) by well-defined statistical benchmark of accuracy, precision, recall and *f*-measure. Based on earlier studies, extant literature and heuristics (Sriram et al. 2010), (Castillo, Mendoza and Poblete 2011), we identify some distinguishing features that can classify a tweet into either "factual-information" or "opinion" category. Noting that a news or factual-information can often influence in generating user-opinions, we treat a tweet as "opinion" in our algorithm, if it contains both facts and personal opinion. Presence of external references like "http", "www" or mention of news media in post or user-description attributes often appears to be a potential indicator of a factual narrative. Formalized presentation, like capitalized first letter in a grammatically correct sentence are also some of the others indicators of a tweet, likely presenting a "factual-information". Simultaneously, presence of emoticons, abbreviation, shortening of words, slang and other Twitter terminologies are stochastically strong indicators of the post being an opinion. Our algorithm also leverages a list of prominent news and institutional agencies, popular social media terminology, common spam words and phrases as our reference data set.

Our proposed algorithm captures real time data from twitter based on a set of keywords. Along with the tweet, it captures all other attributes related to user profile. It scans the tweet message and other attributes like user descriptions, counts of friends and followers, etc. against the identified feature set. The resultant binary values, then goes through a machine learning classifier. We have experimented with eight classifiers and selected maximum entropy classifier, a discriminative classifier that does not assume conditional independence of the feature set. The output labels a tweet into "fact" or "opinion" category.

EXPERIMENT AND EVALUATION

We retrieved around 400k tweets for the keyword "Wells Fargo" using twitter API. After filtering out irrelevant and non-English tweets, we passed a sample data set of 5000 tweets and the rating protocol to two independent coders for manual labeling. The Kappa Statistics is 0.75, indicating substantial inter rater reliability (Landis and Koch 1977). The experiment dataset is a balanced one with 41.76% as facts and 59.24% labelled as opinions. We have experimented with the following eight machine learning classifiers to compare our algorithm with the existing baseline: Naïve Bayes (NB), Maximum Entropy (ME), Support Vector machine (SVM) with linear and non-linear kernel, decision tree classifier, which is a modified version of CART similar to C4.5, Neural Network (NN), Random Forest (RF) and Logistic Regression (LR). We evaluated our algorithm with each of these eight classifiers against the baseline opinion finder (OF), with 10-fold cross validation. Below, we present the output for each standard statistical benchmark namely accuracy, precision, recall and *F*-measure for each iteration for each classifier.

Classifier: Support Vector Machine; Parameter: kernel="linear", C=1 ; 10 fold cross validation											
Scoring Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average and SD
Accuracy	0.7192	0.6696	0.7533	0.7841	0.7004	0.7312	0.7533	0.7433	0.6946	0.6858	0.72(+/-0.07)
Precision	0.6704	0.6021	0.7021	0.7528	0.6710	0.6896	0.7111	0.7108	0.6463	0.6718	0.68(+/-0.08)
F-measure	0.6483	0.5989	0.7021	0.7322	0.6	0.6629	0.6956	0.6704	0.6057	0.5477	0.65(+/-0.11)
Recall	0.6276	0.5957	0.7021	0.7127	0.5425	0.6382	0.6808	0.6344	0.5698	0.4623	0.62(+/-0.15)

Table 1: Scoring Parameter Output for linear Support Vector Machine

Classifier: Support Vector Machine; Parameter: C=1000000.0, gamma="auto", kernel="rbf"; 10 fold cross validation											
Scoring Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average and SD
Accuracy	0.6929	0.6431	0.7489	0.7489	0.7401	0.7224	0.7356	0.7610	0.7079	0.6150	0.71(+/-0.09)
Precision	0.6153	0.5537	0.6697	0.6637	0.6521	0.6347	0.6416	0.6857	0.6238	0.5319	0.63(+/-0.09)
F-measure	0.6464	0.6232	0.7192	0.7246	0.7177	0.6985	0.7196	0.7272	0.6732	0.5347	0.68(+/-0.12)
Recall	0.6808	0.7127	0.7765	0.7978	0.7978	0.7765	0.8191	0.7741	0.7311	0.5376	0.74(+/-0.16)

Table 2: Scoring Parameter Output for non-linear Support Vector Machine

Classifier: K Nearest Neighbor; neighbors =20; 10 fold cross validation											
Scoring Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average and SD
Accuracy	0.7149	0.6475	0.7444	0.7488	0.7136	0.6960	0.7533	0.7566	0.7123	0.6327	0.71(+/-0.08)
Precision	0.6464	0.5603	0.6607	0.6637	0.6	0.5838	0.6637	0.6862	0.6228	0.5471	0.62(+/-0.09)
F-measure	0.6632	0.6190	0.7184	0.7246	0.7280	0.7160	0.7333	0.7179	0.6859	0.5829	0.69(+/-0.10)
Recall	0.6808	0.6914	0.7872	0.7978	0.9255	0.9255	0.8191	0.7526	0.7634	0.6236	0.78(+/-0.19)

Table 3: Scoring Parameter Output for K-Nearest Neighbor

Classifier: Decision Tree; modified Version of CART similar to c4.5										
Scoring Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average and SD
Accuracy	0.6929	0.6475	0.7533	0.7488	0.7400	0.7224	0.7400	0.7610	0.7079	0.6150	0.71(+/-0.09)
Precision	0.6153	0.5603	0.69	0.6637	0.6521	0.6347	0.6446	0.6857	0.6238	0.5319	0.63(+/-0.10)
F-measure	0.6464	0.6190	0.7113	0.7246	0.7177	0.6985	0.7255	0.7272	0.6732	0.5347	0.68(+/-0.12)
Recall	0.6808	0.6914	0.7340	0.7978	0.7978	0.7765	0.8297	0.7741	0.7311	0.5376	0.74(+/-0.16)

Table 4: Scoring Parameter Output for Decision Tree Classifier

Classifier: Random Forest Classifier 10 fold cross validation											
Scoring Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average and SD
Accuracy	0.6929	0.6431	0.7488	0.7488	0.7401	0.7224	0.7356	0.7610	0.7079	0.7345	0.72(+/-0.07)
Precision	0.6153	0.5537	0.6697	0.6637	0.6521	0.6347	0.6416	0.6857	0.6238	0.6701	0.64(+/-0.07)
F-measure	0.6464	0.6190	0.7192	0.7246	0.7211	0.6985	0.7196	0.7272	0.6732	0.5347	0.68(+/-0.12)
Recall	0.6808	0.6914	0.7765	0.7978	0.7978	0.7765	0.8297	0.7741	0.7311	0.4193	0.73(+/-0.22)

Table 5: Scoring Parameter Output for Random Forest Classifier

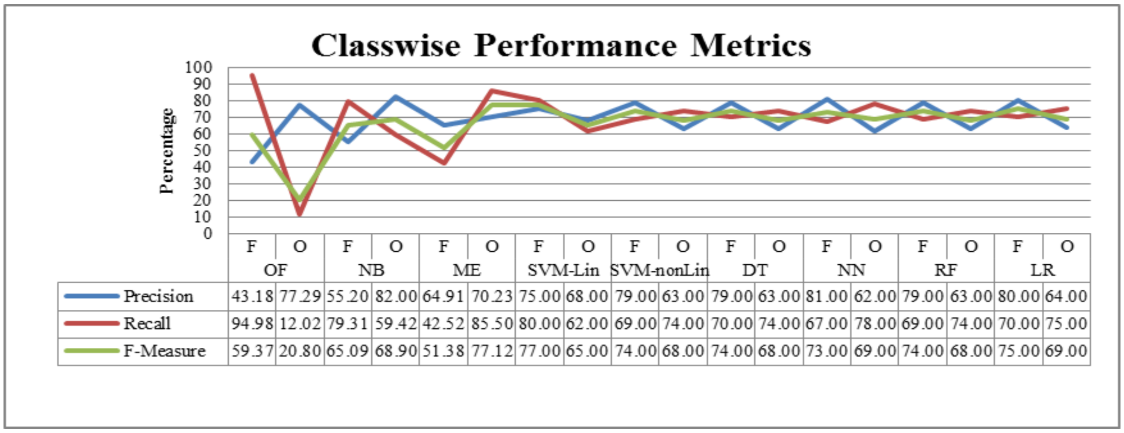
Classifier: Logistic Regression 10 fold cross validation											
Scoring Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average and SD
Accuracy	0.7017	0.6431	0.7488	0.7577	0.7004	0.7224	0.7400	0.7699	0.7079	0.7477	0.72(+/-0.07)
Precision	0.6226	0.5537	0.6697	0.6666	0.6710	0.6302	0.6446	0.6880	0.6097	0.6764	0.64(+/-0.08)
F-measure	0.66	0.6232	0.7192	0.7393	0.6	0.7042	0.7255	0.7425	0.6944	0.7076	0.69(+/-0.09)
Recall	0.7021	0.7127	0.7765	0.8297	0.5425	0.7978	0.8297	0.8064	0.8064	0.7419	0.75(+/-0.17)

Table 6: Scoring Parameter Output for Logistic Regression Classifier

Maximum-Entropy 100 Iteration, MAXENT GIS ;	Naïve Bayes Configuration: 10 Iteration
0.725663716814	0.734513274336
0.66814159292	0.690265486726
0.752212389381	0.747787610619
0.778761061947	0.743362831858
0.761061946903	0.752212389381
0.734513274336	0.690265486726
0.725663716814	0.730088495575
0.769911504425	0.721238938053
0.742222222222	0.702222222222
0.688888888889	0.671111111111
Mean : 73.47040315%	Mean : 71.83067847%

Table 7: Classification Accuracy for Naïve Bayes and Maximum Entropy Classifier

On average, in terms of accuracy measure, our algorithm performs significantly better (72.00%) than opinion finder (46.29%), with highest being observed with maximum entropy classifier (73.47%). Recall (from 48.52% to 70.00%) and *F*-Measure (37.78% to 70.00%) have also shown significant improvement. Precision measure has also shown a moderate improvement (from 62.29% to 72.00%). Figure1 summarizes the performance metrics of precision, recall and F-Measure for each class, namely news or factual-information dented as "F" and opinion, dented as "O".



	OF		NB		ME		SVM-Lin		SVM-nonLin		DT		NN		RF		LR	
	F	O	F	O	F	O	F	O	F	O	F	O	F	O	F	O	F	O
Precision	43.18	77.29	55.20	82.00	64.91	70.23	75.00	68.00	79.00	63.00	79.00	63.00	81.00	62.00	79.00	63.00	80.00	64.00
Recall	94.98	12.02	79.31	59.42	42.52	85.50	80.00	62.00	69.00	74.00	70.00	74.00	67.00	78.00	69.00	74.00	70.00	75.00
F-Measure	59.37	20.80	65.09	68.90	51.38	77.12	77.00	65.00	74.00	68.00	74.00	68.00	73.00	69.00	74.00	68.00	75.00	69.00

Figure 1: Class wise performance metrics comparison between baseline and classifiers

In terms of precision measure for the news or "fact" class, average precision observed in our approach is 74.14 against 43.18 for the baseline - opinion finder. However, in terms of recall measure, baseline (94.98) outperforms our average recall value 68.35 with the highest value (80.00) observed in SVM linear kernel. For both recall and *F*-measure statistics our algorithm demonstrates significant better performance that the baseline. Our average recall and *F*-measure are 72.74 and 69.13 again the baseline value of 12.02 and 20.81 respectively. The most important features that distinguishes a fact from opinion emerged to be the presence of external reference. Mentioning of news media in tweet is another important distinguishing factor that supports our assertion.

CONCLUSSION

In this paper, we conclude that a pre-determined, lexicon corpus based, subjectivity detection method does not perform well in the context of noisy microblog. We proposed an alternative method to classify news from opinion in Twitter. Our algorithm leverages a rich set of metadata from users profile, message propagation and publicly available reference data of social media terminology along with the post. The extracted feature-set from Twitter are experimented with eight machine-learning classifiers, in a supervised setting. The study empirically demonstrates a significant improvement in classification accuracy of our algorithm over the existing baseline. In terms of future research direction, we would like to enhance this algorithm to attain language independence by experimentation with different non-English social media data corpus and associated reference dataset. For practical implication, we may integrate our algorithm with other opinion-mining applications like election polling predictions or consumer sentiment evaluation where user's opinion are more relevant than factual description of an event. Subsequently online applications that capture breaking news from social media may find our algorithm appropriate to filter out user's opinion to facilitate their focus on captured factual narratives.

REFERENCES

1. Atefeh, F., and Khreich, W. 2013. "A Survey of Techniques for Event Detection in Twitter," Computational Intelligence.
2. Castillo, C., Mendoza, M., and Poblete, B. 2011. "Information Credibility on Twitter," *Proceedings of the 20th international conference on World wide web*: ACM, pp. 675-684.
3. Landis, J. R., and Koch, G. G. 1977. "The Measurement of Observer Agreement for Categorical Data," *biometrics*, pp. 159-174.
4. Liu, B. 2010. "Sentiment Analysis and Subjectivity," *Handbook of natural language processing* (2), pp. 627-666.
5. Riloff, E., Wiebe, J., and Wilson, T. 2003. "Learning Subjective Nouns Using Extraction Pattern Bootstrapping," *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*: Association for Computational Linguistics, pp. 25-32.
6. Song, G., Ye, Y., Du, X., Huang, X., and Bie, S. 2014. "Short Text Classification: A Survey," *Journal of Multimedia* (9:5), pp. 635-643.
7. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. 2010. "Short Text Classification in Twitter to Improve Information Filtering," *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*: ACM, pp. 841-842.
8. Wiebe, J. 2002. "Instructions for Annotating Opinions in Newspaper Articles,").
9. Wiebe, J., and Riloff, E. 2005. "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," *International Conference on Intelligent Text Processing and Computational Linguistics*: Springer, pp. 486-497.
10. Yu, H., and Hatzivassiloglou, V. 2003. "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences," *Proceedings of the 2003 conference on Empirical methods in natural language processing*: Association for Computational Linguistics, pp. 129-136.