Recent developments in Social Media Mining

Peter Wlodarczak, Jeffrey Soar, and Mustafa Ally

***Abstract*—**Mining the Web has become a very active area of research. It has gained in importance not only for academia but also for companies since users use the Web to share opinions, ideas and concerns on many subjects and extracting this information can reveal peoples motivations, behavioral patterns and intents. Companies have realized the potential of Web mining for detecting new trends, for getting user opinions on their products and services and for finding new business opportunities. In recent years Social Media such as Facebook and Twitter has been used since it facilitated the publishing and distribution of user generated content. For the first time in history we have now an unprecedented amount of opinionated data that is publicly accessible from anywhere in the world. Analyzing Social Media has been used among other for targeted marketing, opinion mining and has the potential to change the way companies do business. However, Social Media mining is challenging since it has to extract contextual information from human generated content. It has to analyze the meaning, syntax, sentiment polarity of natural language. Social Media mining is in the area of cognitive computing. Cognitive computing systems differ from traditional computing systems since they cannot use preconfigured rules and programs. That is why often Machine Learning techniques are used for Social Media analysis. They have the capability to learn from data and adapt to new problems and domains. This paper describes the state of the art techniques that have been used in recent research and proposes and approach for Social Media mining based on Machine Learning methods.

***Keywords*—**Machine learning, opinion mining, predictive analytics, Social Media Mining, natural language processing, cognitive computing

I. Introduction

SINCE the advent of Web 2.0 technologies, the Web has seen a shift from publisher created to user created content [1]. Web 2.0 and *Social Media* (SM) facilitated the publishing of content by omitting the need to be able to program. Everyone can now post opinions, views, ideas and interests on any topic and they are accessible in real time from anywhere in the world. Facebooks' data volume grows by more than 500 TB every day [2]. On Twitter, more than 500 million Tweets are sent per day by Twitters own account [3]. This resulted in an unprecedented amount of opinionated data globally accessible for anyone from anywhere. Not surprisingly analyzing SM data has become a very active area of research since mining people's opinions can reveal relevant market research information that result in more targeted business decisions. SM analysis has also been used for making predictions on the development of financial markets [4], box office sales [5] or disease outbreaks [6] to name a few. SM mining is a form of Web mining. Web mining, the term is defined as extract needed information from users from the Web [30]. To effectively analyze the large volumes of data, Big Data techniques have to be applied. First the SM data has to be analyzed for its opinion polarity. Opinion mining techniques are often adopted using Machine Learning (ML) techniques. ML is a growing area of data analysis. ML schemes are trained using historic data mimicking human learning. Once trained the ML scheme is applied to new, unseen data to make predictions. For instance, an ML algorithm can learn from past customers who switched to a new company to predict which customers are likely to change in the future.

Opinion mining, also called *sentiment analysis*, is a subarea of *natural language processing* (NLP). It analyses people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [21]. The emerging research area of opinion mining deals with computational methods in order to find, extract and systematically analyze people's opinions, attitudes and emotions towards certain topics [7]. Its aim is to *classify* documents, SM posts, according to their *sentiment polarity*. The classification can be *binary*, for instance positive or negative user reviews, or *multiclass*, where posts such as Tweets are divided according to mood states such as "excited", "skeptical" or "angry". Opinions can be expressed at the *entity level*, a product as a whole, for instance "the new Tesla is excellent", or at the *aspect level*, for instance "the voice quality of the new iPhone is good but battery life time is short", where individual features of an entity are evaluated. Opinion mining on SM has not only been used in academia, there is a growing interest from the industry to find out what users think of their products or services, to detect trends and find new business opportunities. In the past companies had to conduct surveys to collect and assess customer satisfaction. Using SM, there is no need to issue questionnaires to a sample set of users since all data can be analyzed. This process is also called *SM listening*.

ML is an area of *Artificial Intelligence* (AI). ML techniques detect patterns in data and can adapt when exposed to new data. For instance spam filters often use ML algorithms since

P. Wlodarczak is a research student at the Faculty of Business, Education, Law and Arts, University of South Queensland, Australia (corresponding author phone: 076-488-5774; e-mail: wlodarczak@ gmail.com).

J. Soar is a professor at the Faculty of Business, Education, Law and Arts, University of South Queensland, Australia (e-mail: Jeffrey.soar@usq.edu.au).

M. Ally is a lecturer in Information Systems at the School of Management and Enterprise at the University of South Queensland, Australia (e-mail: allym@usq.edu.au).

they can adapt when new types of spam appear. Opinion mining combined with ML techniques has been used in many domains. A prominent success story was the football finals in Brazil, where Google correctly predicted the winner of 11 out of the 12 final games using ML techniques [13]. Microsoft's Cortana even correctly predicted the winners of all finals [14], however less is known about their predictive model.

ML is a well-studied area, and ML techniques have been applied to many Big Data analysis problems. However applied to SM there are challenges due to the large volumes and the variety of the data and due to the peculiarities of SM such as slang and jargon used in posts. Also natural language is difficult to analyze automatically since there is no ground truth. Word sense disambiguation, lemmatization, sarcasm detection and sentiment holder detection remain challenging tasks. For instance the word "meeting" can be a verb, then the word stem is "to meet", or it can be a noun, then "meeting" is already the word stem. This paper describes the state-of-the-art Big Data analysis techniques that have been adopted in recent studies to mine opinions in SM and make predictions based on historic SM data. It proposes a four phase approach for collecting and analyzing SM data and to make predictions based on ensemble learning.

II. Previous work

SM analysis has been used in many domains. Sentiment analysis is a growing area of SM mining. Nowadays social media services such as Twitter and Facebook are increasingly used by online users to share and exchange opinions, providing rich resources to understand public opinions [15]. *Social correlation theories* have been proposed for sentiment analysis by some authors [15]. Other studies have used computational approaches for opinion mining. Different opinion mining algorithms have been analyzed and investigated for their effectiveness [7]. Sentence splitting, stemming, part of speech tagger and parsing algorithms were applied. The researchers concluded that extensive text preprocessing and using algorithms that can effectively process noisy content performed best. Machine learning (ML) techniques such as *supervised methods* based on *naïve Bayesian* and *Support Vector Machine* classification as well as *unsupervised methods* using part of speech tagging have been proposed for political opinion mining on SM [16]. ML techniques have also been used for target oriented opinion mining using a *bag-of-words* supervised classifier [17]. The researchers achieved a classification accuracy of 0.69 for classifying Tweets. Other approaches used *Latent Dirichlet Allocation* (LDA) [18]. LDA is a form of *Latent Semantic indexing*. Latent Semantic Analysis (LSA) is a mathematical technique that is used to capture the semantic structure of documents based on correlations among textual elements within them [31]. LDA characterizes every document by a *Dirichlet distribution*. The similarity between documents is then calculated using a distance measure. The authors concluded that the best results were achieved using a *Jaccard index*.

A very active area of research is *predictive analysis* using SM data. Twitter Tweets have been analyzed to make predictions of financial indicators based on public mood states [4]. The authors investigated if there is a *correlation* between certain public moods on Twitter and the development for the Dow Jones Industrial Average (DJIA) using time series analysis. They concluded that certain mood states do correlate with the development of the DJIA. Other studies analyzed whether box-office revenue could be predicted [5]. They concluded that there is a correlation between the number of positive Tweets and box-office revenue. They also found a correlation between the number of Tweets about a movie and the number of spectators. Similar results for stock price and movie box office revenue were obtained by other studies [12] correlating Twitter based time series.

Sentiments can be expressed with emoticons, they have been used for sentiment analysis [8]. Emoticons have been treated similarly to sentiment words to determine the sentiment polarity of SM posts replacing facial expression in person to person interaction.

An important step in SM analysis is *data pre-processing*. Bitter experience shows that real data is often disappointingly low in quality [9]. Text quality can have a significant impact on the opinion mining process and has been analyzed for several algorithms [7]. Several studies developed improved techniques for purifying SM data from noise and irrelevant content. LDA has been used for relevance filtering [12]. LDA is based on *Latent Semantic Indexing* [11]. It creates a latent description of relevant posts that is used to filter out irrelevant content. This paper builds on previous studies and proposes a methodology described in the next chapter.

III. Research methodology

Data mining is a multidisciplinary field in the areas of statistics, databases, machine learning, artificial intelligence, information retrieval and visualization [32]. SM mining encompasses four phases, a data collection phase, a data pre-processing phase, a data mining phase and a post-processing phase. The first two phases comprise the data conditioning tasks where the data is collected and preprocessed for analysis. In the analysis phase, the data is mined for actionable patterns and correlations are searched for [11]. In the post-processing phase, the data is often visualized, or reports are generated. In this phase sometimes predictive analysis is performed, it is also called the predictive phase. It is executed when data is not only mined to understand the underlying structure and detect patterns, but when projections of future events are seeked for, for instance predicted sales volumes of a new product. Each phase can go through several iterations. Data mining typically goes through many iterations until satisfactory results are achieved. The four phases are described in the next chapters.

A. Data collection

SM data can be accessed through *Application Programming Interfaces* (API). However some SM sites don't offer APIs or

have closed them, for instance LinkedIn has almost entirely removed API access to its data. Some of the big SM sites such as Facebook or Google+ offer a whole set of APIs for data access. The data can thus be collected programmatically using Java, Python or any other programming or script language. For instance, Facebook has a Graph API that can be used for posting and retrieving data. Twitter has a query API to access historic tweets. Twitter also provides a streaming API to access real-time data. The "firehose" API gives access to 100%, the "gardenhose" API to 10% and the "spritzer" API to 1% of real-time Tweets. Recently Facebook has also added a streaming API to its interfaces. However on many SM sites free access is usually limited. Full access such as Twitters "firehose" API is very costly, only the "gardenhose" and "spritzer" API are free. Also, SM sites have often changed access to their data through APIs for instance by introducing quotas. Facebooks streaming API, the Public Feed API, is restricted to a limited group of researchers and one cannot apply for it.

If no API is present, screen scrapers can be used. Screen scraping is the process of extracting human readable content from another program. For SM data Web scraping, also called Web harvesting, can be used. Web scrapers are often browser plugins, for instance NVivo's NCapture (http://www.qsrinternational.com) plugin. They hook into the browsers Document Object Model (DOM) and parse Web pages into d DOM tree.

A third possibility is using third party Web tools such as Topsy (http://topsy.com) or Gnip (https://gnip.com). However they are usually not free of charge and access to premium features is costly.

A data mining task usually begins with understanding the domain. Opinions are expressed differently depending on if they are about political events, products or holiday destinations. So in the data collection phase not only the access methods have to be evaluated, but also the search terms have to be defined. For instance Twitter has a powerful query interface which allows to include and exclude search terms, to give the time range for the query, to do conditional searches and even the define the attitude of a Tweet. For instance the query:

```
"iPhone S6" -Apple since:2015-07-19
+exclude:retweets :)
```

Finds all Tweets containing the exact match "iPhone S6", not containing the word "Apple" since 2015.7.19, excluding retweets with a positive attitude.

B. Data pre-processing

Raw data is seldom in a form that is useful for data mining. SM data is noisy, full of irrelevant information for analysis and contains a lot of spam. The data has thus to be cleaned, and *relevance filtered* first. Data cleaning is a time-consuming and labor-intensive procedure, but one that is absolutely necessary for successful data mining [9]. For opinion mining, only the phrases expressing the sentiment have to be extracted. Opinion mining is highly domain specific, and the first task is to define the sentiment words to look for. For instance an opinion can be expressed using sentiment word such as "great", "excellent", "awful", using verbs such as "like", "love", for instance "the new iPhone is great" or "I like this car". Sentiments can also be expressed using idioms such as "this car cost me an arm and a leg" or words that don't hold a sentiment, for instance "this beer is flat". Other common tasks in opinion mining are stop words removal, finding word stems using *stemming algorithms* and grouping the different inflected forms of a word so it can be analyzed as a single item using *lemmatization algorithms*. Once the sentiment words or phrases have been defined for a specific domain, the SM posts can be analyzed for their *sentiment polarity*. A list of sentiment words is called a *sentiment lexicon*, and these approaches are called sentiment lexicon based opinion mining. Other popular approaches use bag-of-words or part-of-speech tagging.

ML algorithms usually don't process text as input, they need a *feature vector*. Texts have to be represented in the vector space based on *Vector Space Modeling* (VSM). This process is called feature extraction. Feature vectors can be word frequencies of sentiment words, *part of speech* (POS) tags, or sentiment polarity shifters, or *word weights*. *Term Frequency* and *Inverse Document Frequency* (TF-IDF) is one of the best known term weighting methods [23]. It is defined as:

$$w_{t,d} = tf_{t,d} \times \log(\frac{N}{df_t}) \quad (1)$$

where $tf_{t,d}$ is the number of occurrences of term t in the document d, N is the number of document in the collection and df_t, is the number of documents, in which term t appears [23]. The posts are then classified according to their sentiment polarity based on their similarity using a distance measure such as the Euclidian distance, the Manhattan distance or the Chebyshev distance. Very good results have also been achieved using the Bhattacharyya distance [33].

Another approach to creating inputs for ML algorithms is creating *bag-of-words*. A bag-of-words is a list of all the words in a text disregarding grammar or word order. They are often used when mining news articles for opinions, but can be applied to SM data too. Bag-of-words based approaches model news articles by vector space model which translates each news piece into a vector of word statistical measurements, such as the number of occurrences, etc. [22]. Bag-of-words are suitable as inputs for ML algorithms. They have the advantage that some of the data cleaning steps such as stemming or lemmatization can be omitted, however, they tend to perform less well when a lot of slang terms or special characters such as emoticons are used in posts.

C. Data mining

Data is mined to understand the underlying structure of the data and to make predictions based on historic data. It is the process of finding useful, actionable patterns in data and transform the raw data into knowledge. Opinion mining of SM

posts is a *text classification* problem where posts are classified according to their sentiment polarity. SM posts can also be categorized using *clustering techniques* [24]. ML techniques are a suitable way for classification as well as clustering.

There are many ML learning techniques. They fall into three categories, *supervised*, *unsupervised* and *semi-supervised* ML models. Supervised ML techniques are used when the class label is known. For instance, when classifying Tweets into positive and negative Tweets, the class labels are "positive" and "negative". Supervised techniques are used for classification and regression, unsupervised techniques are used for clustering when the class label is not known. Semi-supervised methods are used when there is a small amount of labeled data and large amounts of unlabeled data. For instance in genome sequencing there is usually a small sample size *n* and a large number of markers *p*, "large *p* small *n* problem". Semi-supervised techniques can alleviate this problem [20]. The model is first trained using the small sample set, then it is applied to the large, unlabeled data set.

Ultimately we want to find a decision function f, which classifies SM posts according to their sentiment polarity. In the case of binary sentiment classification, we group posts into positive, *P*, and negative, *N*, reviews. If we denote the set of all posts by T, we search for a function $f:T \rightarrow \{P,N\}$. We use a random set of pre-classified training posts $\{(t_1,c_1), (t_2,c_2),\ldots, (t_n,c_n)\}$, where $t_i \in T$ and $c_i \in \{P,N\}$ to train the learning scheme.

Experience shows that no single machine learning scheme is appropriate to all data mining problem [9]. Usually, several ML schemes are trained, and the one that has the best classification accuracy will be chosen. ML techniques include *naïve Bayes classifier*, *decision tree induction*, *Support Vector Machines* (SVM), *artificial Neural Networks* (aNN) and *k-Nearest Neighbor* (k-NN), but there are many more. They are well studied and have been applied in virtually any data mining domain. ML techniques such as aNN can handle very complex problems and give good approximations. However, they also tend to become complex themselves making it difficult to optimize. SVM use similar concepts to the perceptron used in some aNN, but they are simpler and tend to have a better classification performance.

1) Support Vector Machines

Support Vector Machines (SVM) are based on *statistical learning theory*. SVM create a feature space or vector space defined by a *similarity matrix* (kernel) and create a hyperplane, an *affine decision surface*, to separate the training set. Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible [9]. They maximize the distance from the closest training samples and transcend the limitations of linear separations by including nonlinear terms and thus creating higher order decision boundaries. The techniques are related to the perceptron, which separates the training data set using a linear function. Perceptrons can be organized in interconnected layers creating a multilayer perceptron, an artificial neural network, to create a nonlinear decision boundary. Multilayer perceptrons allow getting good approximations for very complex, non-linear problems, however, they are complex in itself and they don't learn the maximum-margin hyperplane. SVM are a much simpler alternative and have become very popular in recent research.

If the training data is linearly separable, then a pair (w, b) exists such that:

$w^T x_i + b \geq 1$, for all $x_i \in P$

$w^T x_i + b \leq -1$, for all $x_i \in N$

with the decision rule given by:

$$\int_{w,b}(x) = \operatorname{sgn}(w^T x + b) \qquad (2)$$

where w is termed the weight vector and b the bias (or $-b$ is termed the threshold) [20].

SVM have been used primarily for classification, but they can also be used for regression.

1) Ensemble learning

Combining the output of several different models can make decisions more reliable. This process is called *ensemble learning*. Prominent methods include *bagging*, *boosting* and *stacking*. By combining several weak learning schemes, it is often possible to create a strong one. Ensemble learners have performed astonishingly well, but researchers have been struggling to explain why. For example, whereas human committees rarely benefit from noisy distractions, shaking up bagging by adding random variants of classifiers can improve performance [9]. Ensemble learning can comprise hundreds of models which makes it difficult to understand which factors improve the performance.

Probably the best performing ensemble learning scheme is *boosting* [9]. Boosting combines models that complement each other. The models are of similar type, for instance, decision trees. Boosting iteratively builds models based on the performance of the last model such that the new model is trained on instances that were incorrectly classified by the last trained model. This only works well if each model correctly classifies a significant amount of data. Also boosting doesn't tread models equally but contrary to bagging weights a model's contribution by its confidence.

A boosting method designed specifically for classification is *AdaBoost*. AdaBoost calculates the weight of a model based on the models overall error *e*. The error rate is just the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier [9]. The weight w is then calculated as:

$$w = -\log\frac{e}{1-e} \qquad (3)$$

Ensemble learners have many properties that make them very suitable for SM data analysis. For instance models that identify spam with high accuracy such as the naïve Bayes classifier or perceptron [27] can be combined with models that are performing well in relevance filtering or classification thus

creating a stronger learner than a single trained model.

Ensemble learners adopt a divide and conquer strategy in that they combine different learners with different accuracies in order to obtain a composite model that leverages the weakness of each single model. For example, *Instance Selection* (IS) is often used to handle noise [25]. Combining such a model with a model that is suitable for a specific classification problem can improve classification accuracy and also reduce the effort that goes into data pre-processing. SM data can thus be processed by different models, models that eliminate spam, models for relevance filtering and finally models for the actual classification.

Ensemble learners can handle very complex data mining problems, but they can become very complex themselves which runs counter to *Occam's razor*, which advocates simplicity. Loss of interpretability is a drawback when applying ensemble learning, but there are ways to derive intelligible structured descriptions based on what these methods learn [9]. Ideally instead of having an ensemble of learners, which makes it very difficult to interpret what kind of information has been extracted from what data, a single model would be preferred. If the ensemble learner is composed of decision trees, it is possible to combine them into a single structure, but it might still be difficult to interpret. An alternative are *LogitBoost trees*, which induce trees using *linear-logistic regression models* at the leaves. LogitBoost is an extension to the AdaBoost algorithm. It replaces the exponential loss of Adaboost algorithm to *conditional Bernoulli likelihood loss* [28]. If the LogitBoost algorithm is run until convergence, the result is a *maximum-likelihood, multiple-logistic regression model*. Running till convergence occurs is often not feasible due to performance issues when run against future, unseen data. However, it usually not necessary to wait until convergence to obtain good results. AdaBoost and LogitBoost are a very efficient classification method on *balanced data sets*. In real-world data, it is quite common to have *unbalanced classification data* and extensions to LogitBoost have been proposed [28],[29] to overcome this problem.

1) Deep learning

In recent studies, *Deep Learning* (DL) has been used in many studies with very encouraging results [34-36]. It has been successfully applied on problems that have resisted the best attempts of AI research such image and speech recognition. DL has produced very promising results for various tasks in natural language understanding such as question answering, classification, sentiment analysis and language translation. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level [37]. Similar to ensemble learners, DL consist of stacks of simpler models. Each layer transforms the input into a higher level representation, creating an abstraction of higher order. The learned high-level representations have shown to give state-of-the-art results in areas such as visual data mining and NLP. Probably the most useful property that distinguishes DL schemes from traditional learners such as SVM is that they can extract features automatically. They can be feed with raw data and discover automatically the representations needed for classification. A few notable examples of such models include Deep Belief Networks, Deep Boltzmann Machines, Deep Autoencoders, and sparse coding-based methods [38]. Sometimes they were criticized as a rebranding of aNN. However, we expect to see more successes in DL in the near future since they can take advantage of the large amounts of computational data and need very little engineering by hand.

IV. CHALLENGES

Opinion mining remains a challenging area of research. Next to the regular challenges such as *sentence boundary disambiguation*, *word disambiguation*, and *sarcasm detection*, SM sites have certain properties which pose additional problems.

Spam has become a major issue on the Internet. Fake opinions are very difficult to detect, and opinion spammers often have fake identities (sock puppet, catfish).

Slang and jargon used in SM posts pose a major challenge for opinion mining. It is often specific to certain types of sites such as dating sites, political discourse forums or product review sites. Also, many SM sites have specific characteristics such as the dollar sign denoting a company, e. g. "$AAPL" for Apple Inc. or the hash tag "#" denoting the subject in Tweets. Abbreviations such as LOL (Lough out loud), IMHO (In my humble opinion) or AFAIK (As far as I know) are also widely in use, especially on microblogging sites where the number of characters per post is limited.

Noisy texts pose additional challenges since many ML algorithms such as naïve Bayes don't handle it very well. Also, SM posts tend to be grammatically less correct and have many spelling errors which makes for instance sentiment lexicon based opinion mining or POS tagging less accurate. Often spelling errors are intended, for example for emphasis, e. g. "Goooooooood camera".

Most learning algorithms try to learn from noisy data by modeling the maximum likelihood output or least squared error, assuming that noise effects average out [26]. However, this method only works well for *symmetrical noise distributions*. Sources of noise in SM are typically asymmetrical, and many classification schemes such as naïve Bayes do not work well in these conditions.

SM site users decide themselves if they want to post an opinion on a certain subject, and the *self-selection bias* applies.

Identifying background topics that have been discussed for a long time and that are irrelevant to the public's opinion is another issue that has to be addressed. Text clustering and summarization techniques are not appropriate for this task since they will discover all topics in a text collection [10].

Deep learners have shown very promising results especially in the areas of object recognition and language perception. However, most research was based on supervised learning. Human learning in contrast is mostly unsupervised. Humans and animals discover the world just by observing it, not by labeling it. More research in unsupervised DL is not only desirable, it is to be expected that it will make DL even more useful.

Lastly, there are challenges inherent in ML techniques. Some models such as decision trees or aNNs tend to be overfitted. *Overfitting* occurs when the model becomes too complex and starts to capture noise instead of the actual opinion phrases. It happens when a model becomes too complex, and Occam's razor applies.

V. Conclusions

Ensemble learners have worked surprisingly well when analyzing SM data. They are very robust also when data is noisy. However applying them requires a lot of experience and more research in certain areas is highly desirable. Making ensemble learners simpler by analyzing which features contribute to what extent to the result is one of the goals of our research. Ultimately we would like to have a learner that consists of only one model or at least only a few models with a clear separation of which model extracts what information. Simplifying models without losing predictive performance is an area where we would like to see more research effort.

Data pre-processing is an important step, and there seems to be much less research in data cleaning and feature selection than in the actual data analysis tasks. Spam or fake opinion detection remains difficult and more studies in this area could improve classification accuracy a lot. Feature selection is at least as important as selecting the most suitable learning scheme, and more research could lead to improved data mining results.

Correlation doesn't mean causation. If there is a correlation for instance between the number of positive Tweets and the sales volume of a product it doesn't mean there is also a causal link. It is generally difficult to find the exact causes of sentiment variations since they may involve complicated internal and external factors [10]. A more holistic research approach could analyze the factors that influence positive reviews and product sales and lead to a clearer understanding of the causation.

Most studies treat every post equally. But some posts might be more influential because more people read them, or the poster has a higher authority. There has been some research on finding influential people in SM or in analyzing the online authority of users. Analyzing the impact of for instance opinion Tweets would improve opinion mining since some Tweets might be more influential because they have more followers or are more authoritative. SM posts could then be graded by their influence that would improve the predictive power of SM analysis.

References

[1] P. Wlodarczak, J. Soar, and M. Ally, "Big Personal Data", Social Science Research Network, 2014.

[2] S. McGlaun, "Facebook data grows by over 500 TB daily", SlashGear, 2012.

[3] Twitter (2015, January), About, Available: https://about.twitter.com/company.

[4] J. Bollen, H. Mao, and X.-J. Zeng, Twitter mood predicts the stock market, Journal of Computational Science, vol. 2, pp. 8, 2010.

[5] S. Asur, and B. A. Huberman, Predicting the Future with Social Media, presented at the IEEE Int. Conf. Web Intelligence, 2010, pp. 492-499.

[6] H. Achrekar, A. Gandhe, R. Lazarus, Y. Ssu-Hsin, and L. Benyuan, Predicting Flu Trends using Twitter data, presented at the IEEE Computer Communications Workshops (INFOCOM WKSHPS), 2011, pp. 702-707.

[7] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteský, and A. Holzinger, "Computational approaches for mining user's opinions on the Web 2.0," Information Processing & Management, vol. 50, no. 6, pp. 899-908, 11//, 2014.

[8] N. Oliveira, P. Cortez, and N. Areal, "Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter," in Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, Madrid, Spain, 2013, pp. 1-8.

[9] I. H. Witten, E. Frank, and M. A. Hall, Data Mining, 3 ed., Burlington, MA, USA: Elsevier, 2011.

[10] T. Shulong, L. Yang, S. Huan, G. Ziyu, Y. Xifeng, B. Jiajun, C. Chun, and H. Xiaofei, "Interpreting the Public Sentiment Variations on Twitter," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no. 5, pp. 1158-1170, 2014.

[11] P. Wlodarczak, J. Soar, and M. Ally, "What the future holds for Social Media data analysis," World Academy of Science, Engineering and Technology, vol. 9, no. 1, pp. 545, 2015.

[12] M. Arias, A. Arratia, and R. Xuriguera, "Forecasting with twitter data," ACM Trans. Intell. Syst. Technol., vol. 5, no. 1, pp. 1-24, 2014.

[13] B. Bechtolsheim, (2014, July) Google Cloud Platform is 11 for 12 in World Cup predictions. Google Cloud Platform. [Online]. Available: http://googlecloudplatform.blogspot.ch/2014/07/google-cloud-platform-is-11-for-12-in-World-Cup-predictions.html.

[14] V. Shet, (2014, July) Microsoft's Cortana predicts that Germany will win the FIFA World Cup 2014, sporstskeeda. [Online]. Available: http://www.sportskeeda.com/football/microsofts-cortana-predicts-germany-will-win-fifa-world-cup-2014.

[15] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: a survey," SIGKDD Explor. Newsl., vol. 15, no. 2, pp. 20-29, 2014.

[16] S. Stieglitz, and L. Dang-Xuan, "Social media and political communication: a social media analytics framework," Social Network Analysis and Mining, vol. 3, no. 4, pp. 1277-1291, 2013/12/01, 2013.

[17] V. Hangya, and R. Farkas. Target-oriented opinion mining from tweets. in Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on. 2013.

[18] D. Zlacky, J. Stas, J. Juhar, and A. Cizmar, "Text Categorization with Latent Dirichlet Allocation," Journal of electrical and electronics engineering, vol. 7, pp. 161-164, 05/01, 2014.

[19] A. Kyriakopoulou, and T. Kalamboukis. Text classification using clustering. in Proceedings of The 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Berlin, Germany. 2006.

[20] Yip, K., C. Cheng, and M. Gerstein, Machine learning and genome annotation: a match meant to be? Genome Biology, 2013. 14(5): p. 205.

[21] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2 ed., Heidelberg: Springer, 2011.

[22] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," Knowledge-Based Systems, no. 0, 2014.

[23] V. Hangya, and R. Farkas, "Target-oriented opinion mining from tweets." pp. 251 -254.

[24] B. Liu, Sentiment Analysis and Opinion Mining: Morgan & Claypool, 2012.

[25] S. B. Kotsiantis, "Supervised Machine Learning," Informatica, vol. 31, pp. 19, 2007.
[26] M. D. Schmidt, and H. Lipson, "Learning noise," in Proceedings of the 9th annual conference on Genetic and evolutionary computation, London, England, 2007, pp. 1680-1685.
[27] K. Tretyakov, "Machine Learning Techniques in Spam Filtering," Data Mining Problem-oriented Seminar, U. o. T. Institute of Computer Science, ed., 2004, p. 19.
[28] S. Jie, L. Xiaoling, L. Miao, and W. Xizhi, "A new LogitBoost algorithm for multiclass unbalanced data classification." pp. 974-977.
[29] J. Song, X. Lu, and X. Wu, "An Improved AdaBoost Algorithm for Unbalanced Classification Data ". pp. 109 - 113.
[30] M. Santhanakumar, and C. Christopher Columbus, "Web Usage Based Analysis of Web Pages Using RapidMiner", WSEAS Transactions on Computers, vol. 14, pp. 455-464, 2015.
[31] K. Krishnamurthi, V. Griet, and V. Jntuh. "Capturing the Semantic Structure of Documents Using Summaries in Supplemented Latent Semantic Analysis", WSEAS Transactions on Computers, vol. 14, pp. 314-323, 2015.
[32] P. Wlodarczak, J. Soar, and M. Ally, "Data Process and Analysis Technologies of Big Data," Networking for Big Data, Chapman & Hall/CRC Big Data Series, pp. 103-119: Chapman and Hall/CRC, 2015.
[33] M. Hu, S. Zhou, J. Wei, Y. Deng, and W. Qu, „ Change-Point Detection in Multivariate Time-Series Data by Recurrence Plot", WSEAS Transactions on Computers, vol. 13, pp. 592-599, 2014.
[34] X. Zhao, X. Li, and Z. Zhang, "Multimedia Retrieval via Deep Learning to Rank " IEEE Signal Processing Letters, vol. 22, no. 9, pp. 1487 - 1491 2015.
[35] H. Weilong, G. Xinbo, T. Dacheng, and L. Xuelong, "Blind Image Quality Assessment via Deep Learning," Neural Networks and Learning Systems, IEEE Transactions on, vol. 26, no. 6, pp. 1275-1286, 2015.
[36] K. Noda, Y. Yamaguchi, K. Nakadai, H. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," Applied Intelligence, vol. 42, no. 4, pp. 722-737, 2015/06/01, 2015.
[37] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 05/28/print, 2015.
[38] R. Salakhutdinov, "Deep learning," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, New York, USA, 2014, pp. 1973-1973.