Precise Tweet Classification and Sentiment Analysis

Rabia Batool*, Asad Masood Khattak†, Jahanzeb Maqbool‡ and Sungyoung Lee†
*Department of Biomedical Engineering, Kyung Hee University, South Korea, Email: rabia@oslab.khu.ac.kr
‡Dept of Computer Engineering, Ajou University, South Korea, Email: jahanzeb@ajou.ac.kr
†Department of Computer Engineering, Kyung Hee University, South Korea, Email: asad,sylee@oslab.khu.ac.kr

***Abstract*—The rise of social media in couple of years has changed the general perspective of networking, socialization, and personalization. Use of data from social networks for different purposes, such as election prediction, sentimental analysis, marketing, communication, business, and education, is increasing day by day. Precise extraction of valuable information from short text messages posted on social media (Twitter) is a collaborative task. In this paper, we analyze tweets to classify data and sentiments from Twitter more precisely. The information from tweets are extracted using keyword based knowledge extraction. Moreover, the extracted knowledge is further enhanced using domain specific seed based enrichment technique. The proposed methodology facilitates the extraction of keywords, entities, synonyms, and parts of speech from tweets which are then used for tweets classification and sentimental analysis. The proposed system is tested on a collection of 40,000 tweets. The proposed methodology has performed better than the existing system in terms of tweets classification and sentiment analysis. By applying the Knowledge Enhancer and Synonym Binder module on the extracted information we have achieved increase in information gain in a range of 0.1% to 55%. The increase in information gain has enabled our proposed system to better summarize the twitter data for user sentiments regarding a keyword from a particular category.**

I. Introduction

The rise of social media, such as Facebook, MySpace, and Twitter in less than a decade has changed the general perspective of networking, socialization, and personalization [13]. These social networks have significant impact on the daily life of hundreds of millions of users [17]. Social networks can serve as a source to analyze individual interests and their effects on personal life as they provide huge amount of data about person [24]. Among different social networks, Twitter is one of the most popular micro blogging service [20]. Over 465 million twitter accounts in 2012 have generated 175 million tweets per day [25]. People are increasingly using Twitter to share advice, opinions, news, moods, concerns, facts, and rumors.

Hashtags are used in tweets to categorize and also find relevant information. Hashtags allow users to create communities of people interested in the same topic by making it easier for them to find and share related information [8]. Similarly, users can search for specific health conditions and get all the latest health information in the form of tweets [6]. However, at times people do not use specific keywords in tweets for which we might be interested. For example, if a user posts tips to decrease blood glucose level; this is very informative for diabetic patient; however, is not retrieved when searched with keyword *"diabetes"*. Processing of such short text to understand its context and extract the desired knowledge is a challenging task. It needs sequential processing of the text to mine the actual semantics of the contained information. In addition, this can also answer to questions like: what are these tweets about? How are they related? And what sentiments have user expressed about the topic?

In this research paper, we propose a system to process short text and filter them precisely based on the semantics of information contained in them. The proposed system analyzes tweets to find interesting result and integrates sentiments with each individual keyword. Feature extraction is applied to find keywords and sentiments from tweets expressed by user about the keywords. Afterwards semantic based filtering for specific category using seed list (domain specific) is applied that decreases the information loss. In addition, to maximize the information gain the proposed system applies filtering on keywords, verbs, entities and their synonyms extracted from tweet. To verify and validate the working of proposed system, we have tested it with 40000 tweets containing information of categories, such as diabetes, food, diet, medication, education, dengue, parkinsons, and movies. Using the proposed system, first the knowledge generator is applied to classify tweets into different categories . To enrich the information extracted in previous step, knowledge enhancer and synonym binder module is applied to increase the information gain. Overall, significant improvement from 0.1% to 55% has been achieved using the proposed system.

The rest of the paper is organized as follow. Section II discusses the related work closely aligned with our work. Section III describes the proposed system architecture and its components. Section IV shows implementation and results of the proposed system. Section V concludes the work and explore future research directions.

II. Related Work

Twitter draws researchers attention on different issues and has been used for a variety of purposes, such as marketing, communication, business, and education [16]. In this section, we will uncover some of the related research work on twitter data extraction, its meaningful processing,and twitter based developed applications.

Different analysis tools are available to collect twitter data. Archivist [3] is a service that uses the Twitter Search API to find and archive tweets having specific keyword. Grabeeter [5] on the other hand searches and grabs tweets posted by individuals. Users register with Grabeeter which makes their

respective tweets search-able using Grabeeter if they make their account public. Milan et al. [28] collected and parsed tweets about conferences using Twapperkeeper. The extracted data was stored in ontologies like: SIOC, FOAF, and OPO. Garin Kilpatrick [21] introduced list of all twitter tools to collect and analyze Twitter data. He divided all Twitter tools into 53 categories. These tools provides facility in backup tweets, trend analysis, tweets translation, voice tweet, and Twitter statistics.

Fabian Abel et al. [9], [10], [11] collected extracted entities, topics, and hashtags from tweets to build personalized user profile. They also enriched news with tweets to improve the semantic of Twitter activities. Ilknur Celik et al. [15] studied semantic relationship between entities in Twitter to provide a medium where users can easily access relevant content, they are interested in. Mor Naamen et al. [22] studied the users behavior on Twitter. They applied human coding and qualitative analysis of tweets to understand users activities on Twitter. They analyzed that majority of users focus on self(memoformers) while small portion of users share information with others(informers). Milan et al. [28] extracted tweets topics to map tweet talks to conference topic. They enriched tweets information by adding Dbpedia topics using zamanata - an application to extract keywords from text and connect them to related topics in Dbpedia.

Tetsuya Nasukawa et al. [23] used natural language processing techniques to identify sentiment related to particular subject in a document. They used Markov-modal based tagger for recognizing part of speech and then applied statistics based techniques to identify sentiments related to subject in speech. Jeonghee Yi et al. [29] presented a model to extract sentiments about particular subject rather than extracting sentiment of whole document collectively. This system proceeded by extracting topics, then sentiments, and then mixture model to detect relation of topics with sentiments. Whereas, Namrata Godbole et al. [18] introduced a sentiment analysis system for news and blog entities. This system determined the public sentiment on each of the entities in posts and measured how this sentiment varies with time. They used synonyms and antonyms to find path between positive and negative polarity to increase the seed list.

The same way, the extracted tweet information is used by Fabian Abel et al. [10] for personalized news recommendation. They collected tweets of individual user and based on user interest recommended them news article. Whereas Bernard J et al. [20] performed analysis of Twitter as electronic word of mouth in the product marketing domain. They analyzed filtered tweets for frequency, range, timing, content, and customer sentiments. Bharath Sriram et al. [27] proposed an approach to classify tweets into news, opinions, deals, events and private messages with better accuracy . They used eight basic features from tweets. They did not apply noise reduction techniques which may degrade the performance. Jagan Sankaranarayanan et al. [26] introduced TweetStand to classify tweets as news and non-news. Naive Bayes classier was trained on a training corpus of tweets that had already been marked as either news or junk. After filtering news tweets they clustered tweets into different topics. They also extracted geographic content from each tweet, to determine the clusters overall geographic focus. Sarah Zelikovitz et al. [30] described a method for improving the classication of short text by considering secondary corpus of unlabeled but related longer documents with labeled training data . They used four different data sets(Technical paper, News, Web page titles, Companies) to test the system. They showed that their proposed approach reduced the error rates in text classication by using a large body of potentially uncoordinated background knowledge. Somnath et al. [12] and Xia et al. [19] used wordnet and wikipedia to cluster short text precisely.

Unlike standard text with lot of words, tweets which consist of few words brings a great challenge in classification.So classifying and analyzing them for sentiments just on the available keyword is not enough to understand the real semantics of tweets. There is a need to precisely parse and process the tweets for their contained knowledge which is the focus point of this research paper. This will help to classify the tweets appropriately and analyze user sentiments far better for a keyword in a given context.

III. The Proposed System Architecture

In this section, we present our proposed system for Twitter data processing. Our system uses Twitter data and performs parsing, domain specific classification, and sentiment analysis. The proposed system has also found overlap of information in short text by using precise filtering on tweets. To extract tweets, we use Archivist(visitmix.com/work/archivist-desktop/), a service that uses Twitter Search API to find and archive tweets. Table I shows the number of tweets collected for different keywords which are used for the proposed system evaluation. The overall architecture of the proposed system is shown in Figure 1. For extraction of keywords, entities and sentiments we used Alchamy API [1]. Alchemy API utilizes natural language processing technology and machine learning algorithms to analyze content. It can extract keyphrases, named entity, and topic level sentiments. Keyphrases are actually metadata of text returned by Alchemy API. Alchemy API is capable of identifying people, companies, organizations, cities, geographic features, and other typed entities within text. It can extract 28 types of entities from text which contains hundred of further subcategories. A series of statistical algorithms are combined with a huge data-set describing the world's objects, individuals, and locations. It also combine subtypes of entity which provides detailed ontological mappings for an entity, for instance identifying a Person as a Politician or Athlete. The details on component wise processing and analysis of tweet data are explained in the following subsections.

A. Preprocessor

The tweets need to be preprocessed before extracting meaningful information from it.

Tweets returned by Archivist are in XML format that need preprocessing before storing them in a repository. DOM parser is used to parse XML document and store them in a repository (a relational database). Figure 2 shows data returned by Archivist. Parser split data into *Username*, *TweetDate*, *Status*, *TweetID*, and *image* fields.

TABLE I: Number of tweets collected for each keyword using Archivist

Topic	Count
Food	2584
Diet	13983
Blood pressure	3696
Medication	2023
Parkinson	1941
Dengue	2114
Movies	3500
Diabetes	6339
Education	5289

Twitter data also contains slangs and repeated character like *plz* and *gooood* instead of *please* and *good* [14]. Users make use of repetition of same character in tweets to emphasize on a word. This kind of noise can effect knowledge extraction process. For instance, the tweet shown in Figure 2 has positive sentiments about exercise but Alchemy API can not identify positive sentiments due to word *'gooood'*. To remove slangs, the proposed system facilitate in 1300 slangs removal from tweets. Slang remover replaces *'plz'* and *'gooood'* in Figure 2 with *'please'* and *'good'* respectively. Twitter data also contains spelling mistakes; however, spelling corrections are out of the scope of this research paper.

B. Knowledge Generator

The objective is to extract valuable information from tweets and classify the tweets into different categories based on the knowledge contained in them. The collected tweets are given to Alchemy API. It accepts unstructured text, processes it using natural language processing and machine learning techniques, and returns keywords and sentiments of users about keywords. The proposed system extracts participating keywords and their associated sentiments using Alchemy API which is able to extract sentiments at topic level. For instance, Table II shows keywords and associated sentiments, extracted by knowledge

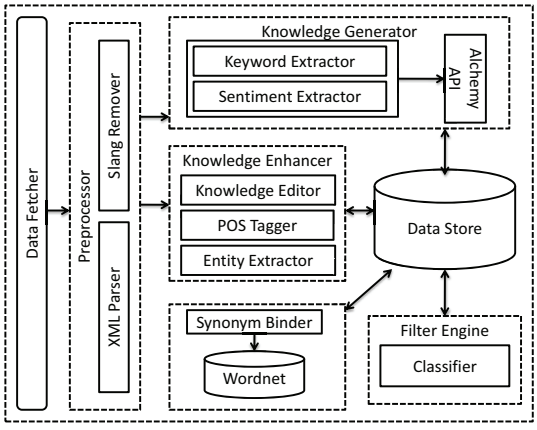


Fig. 1: The proposed system architecture for tweet classification and sentiment analysis

```
<Tweet
 TweetStatus="Unapproved"
 Username="dellimaquen"
 TweetDate="2012-09-12T16:57:47+09:00"
 Status=" I am Scott Malkinson and I have got diabetes plz help
 TweetID="245793316506042368"
Image="http://a0.twimg.com/profile_images/2569014433/IMG00354-20120526-
1717_normal.jpg" />
<Tweet  TweetStatus="Unapproved"
 Username="dellimaquen"
 TweetDate="2012-09-12T16:59:47+09:00"
 Status= Exercise is very gooood for diabetic patient."
 TweetID="245793316506042368"
Image="http://a0.twimg.com/profile_images/2569014433/IMG00354-20120526-
1717_normal.jpg" />
```

Fig. 2: Data collected from Twitter in XML format

TABLE II: Knowledge extracted by knowledge generator

Tweets	Keywords	Sentiments
Exercise is very good for diabetic patient	Exercise	Positive
	Diabetic patient	Neutral
I am Scott Malkinson and I have got diabetes please help	Scott Malkinson	Neutral

generator from tweets shown in Figure 2. After extraction of knowledge, all tweets, participating keywords, and associated sentiments are stored in the repository for further processing as discussed below. However, the information extracted by knowledge generator is of low precision. So, it needs further processing to better classify tweets and analyze sentiments of domain specific keywords.

C. Knowledge Enhancer

Knowledge enhancer module add additional knowledge which was not extracted as keyword by Alchemy API. The proposed system uses part of speech tagging and entity extraction on tweets and add additional data to the knowledge extracted by Alchemy API. Entity extraction using Alchemy API helps in extracting entities, not extracted as keyword. Table III shows example of tweets in which keyword extraction in knowledge generator ignores *'diabetes'* which is very important in tweets classification. *Knowledge enhancer* successfully extracts this information from tweets. To achieve this, the proposed system has incorporated the addition of subjects, verbs, objects, and entities in knowledge; however, just addition of verb and entities increases information collected from tweets.

D. Synonym Binder

Synonym binder is yet another additional step of the proposed system to increase the information gain from the tweets. The proposed system binds synonyms with each entity and keyword extracted by knowledge generator and knowledge enhancer. For example, it binds word *workout* with its synonyms *exercise* and exercise is present in our seed list but workout is not there. It also covers many word structure problems associated with words e.g., it extracts synonym of *calories* as *calorie* and *exercises* as *exercise*. Wordnet dictionary is used to bind synonyms with entity and keywords and let us know

TABLE III: Knowledge generator and knowledge enhancer knowledge

Tweet	Knowledge extracted by knowledge generator	Knowledge extracted by knowledge enhancer
I am Scott Malkinson and I have got diabetes plz help.	Scott Malkinson	Scott Malkinson, diabetes
RT @qytaralore: physiology viagra online cialis commercial ads http://t.co/cPMJqQ6w impotence in young men diabetes.	physiology viagra online, commercial ads, young men, impotence	viagra, diabetes

the real sense of these. Jaws API [7] has been used to get synonyms of words from Wordnet. Synonym binder connects synonyms with words and store them into data store to classify data more precisely.

E. Filter Engine

For classifying tweets into different categories on the basis of knowledge extracted from tweets, the proposed system applies filtering on the extracted knowledge. The filtering process is domain specific. Our focus in this research is on health-care so, we have used medical terms. To classify the tweets, seed list of medical terms is used to identify which category the extracted knowledge belongs to.

Filter engine classifies data on the basis of seed list and stores them in the repository. Use of seed list benefits in such a way that without seed list we just have tweets containing *"diabetes"* keyword in them but with seed list we also filter those tweets which can be useful for diabetic patient with no explicit diabetic keyword in them. For example *"Morning walk is very helpful to maintain Blood glucose"*. This tweets was not filtered when we search Twitter for diabetes; however, the proposed system has successfully classified this tweet as diabetic tweet.

IV. Implementation and Result

We have tested our system for diabetes and have successfully and correctly classified more tweets than the existing systems. Our target implementation in this research is diabetic relevant data. By classifying, system has collected all diabetic information from diabetic related tweets. This precise keyword/seed list based filtered data is used for tweets classification and sentiment analysis. To construct the seed list, we have explored several online resources [2] and [4] to identify diabetes related terms and avoid information loss. Google Refine is used to overcome redundancy problems and formatting. To measure accuracy of our proposed system, we have filtered tweets using seed list for diabetes (discussed later). Moreover, these tweets are also useful for clustering, trend analysis, and recommendations. The detail process for data collection and experiments are given below.

A. Data Collection

Using Archivist, we have collected almost 40,000 tweets for 43 days which are used for system testing, verification, and validation of our claims. Archivist is fed with keywords to crawl twitter and collect tweets. The detail on keywords and number of respective tweets extracted are provided in Table I. For each keyword frequency of tweets is different because number of tweets on specific topic varies for different days, different timings, and user interest.

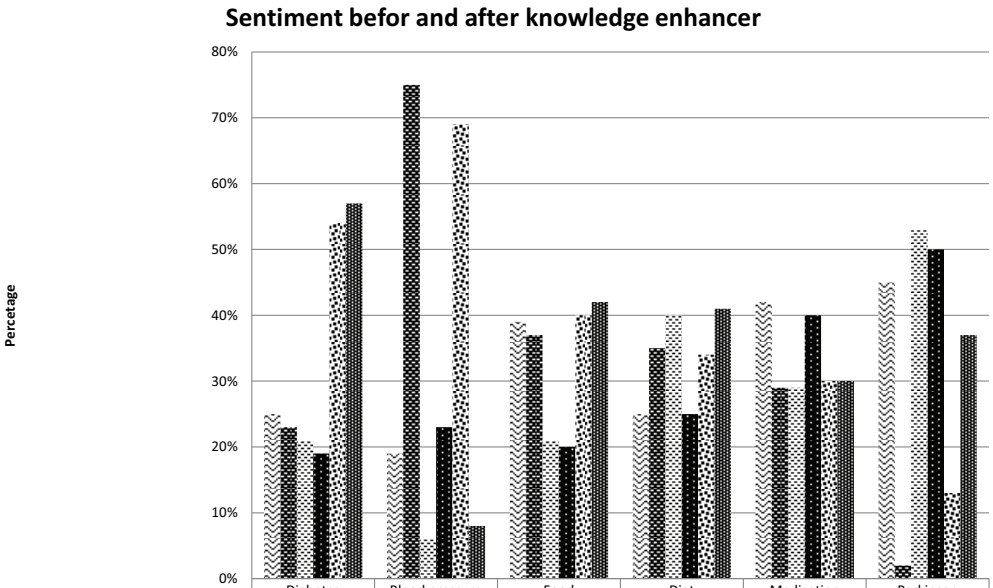
As discussed above, the domain in focus for this research work is diabetes. The seed list consist of diabetic term. It is composed of two different online resources having diabetic terms. After collecting diabetic terms, proposed system splits them into different categories. Online resources from which we collected diabetic terms contains terms and definition. To identify category of term, system has processed definition of terms using natural language processing to categorize them. Diabetic terms are divided into 30 categories. Major categories are diabetic test, condition, body cell, diabetic study, professional, devices, and medicine. For example, proposed system found following definition of *"hyperinsulinemia"*, *"a condition in which the level of insulin in the blood is higher than normal caused by overproduction of insulin by the body"*. By applying natural language processing the proposed system identified that it is a condition and has labeled this term as condition. We could not label 81 diabetic terms in any category so these are kept under *"other"* category. The proposed system is trained with 417 terms which are then used to classify the collected tweets in different categories.

B. Testing

The proposed system has processed 40,000 tweets of different categories for testing and verification. By considering the keywords returned by Alchemy API, 3874 diabetic tweets were classified from all categories. However, when the proposed knowledge enhancer and synonym binder in addition to knowledge generator are applied then the proposed system has classified 8636 diabetic tweets from all categories because knowledge enhancer extract entities from tweets which are important for categorization. Its accuracy has increased in a range of 0.1% to 55% for different categories as shown in Table IV. Information gain is due to verb, entities and their synonyms in tweets which are extracted by the proposed system; however, missed by Alchemy API as keywords. To increase this information gain, our system has collected verb and entities from tweets, bind synonyms with entities and

TABLE IV: Information dispersion for diabetes

Keyword	Diabetes tweet detected without knowledge enhancer	Diabetes tweets detected after knowledge enhancer	Diabetes tweets detected after synonym binder
Diabetes	34%	89.4%	89.5%
Food	0.9%	1.2%	1.7%
Diet	0.8%	1.1%	1.8%
Blood pressure	44.3%	44.8%	72.8%
Medication	0.3%	0.3%	1.2%
Parkinson	0.4%	0.4%	0.4%
Dengue	0%	0%	0.1%
Movies	0%	0%	0.1%
Education	0.3%	1.0%	1.1%



	Diabetes	Blood pressure	Food	Diet	Medication	Parkinson
Positive befor knowledge enhancer	25%	19%	39%	25%	42%	45%
Positive after knowledge enhancer	23%	75%	37%	35%	29%	2%
Negative befor knowledge enhancer	21%	6%	21%	40%	29%	53%
Negative after knowledge enhancer	19%	23%	20%	25%	40%	50%
Neutral befor knowledge enhancer	54%	69%	40%	34%	30%	13%
Neutral after knowledge enhancer	57%	8%	42%	41%	30%	37%

Fig. 3: Sentiments division before and after knowledge enhancer

keywords and have applied filtering by addition of verb and entities with keywords and their synonyms. Table III shows the information ignored by the knowledge generator is extracted by knowledge enhancer module. Table IV shows the collective results of knowledge generator, knowledge enhancer and synonym binder. The results show visible increase in information gain of knowledge enhancer and synonym binder. With the help of knowledge enhancer and synonym binder, we are able to classify 89% of diabetic tweets correctly whereas, knowledge generator has classify 34% of diabetic tweets. Remaining 11% diabetic tweets are ignored due some spelling mistakes and language other then English because system works just for English. The proposed system has also increased the number of filtered tweets for other categories.

The proposed system has also highlighted the chances of information loss or misleading information from tweets if it has only been filtered by keywords. When data searched from Twitter using keywords, it does not return all related data which can provide useful information. Using short text classification, we try to extract all information related to specific topic. Table IV shows information exist in data collected using other keywords. We have tested proposed system by filtering tweets for diabetes from all tweets collected using different keywords and we found overlap in these tweets shown in Table IV. Results show that each category contains information about diabetes except *Movies* and *Dengue* in knowledge generator and knowledge enhancer, and *Movies* in synonym binder. Table IV shows that 1.7% tweets from food and 1.8% tweets from diet contains valuable information about diabetes. Information diffusion varies in each category but for blood pressure we found 72.8% of tweets which can give information about diabetes too; however, are ignored in the existing methodologies.

The proposed system has also analyzed tweets for sentiment analysis. As knowledge enhancer increases information gain, it also increases the result maturity for sentiments divisions. Figure 3 shows division of positive, negative, and neutral sentiments before and after applying knowledge enhancer. Knowledge generator only using Alchemy API ignores entities and verbs from tweets. So, when we make division of positive, negative, and neutral sentiments, we got different division. However, with the addition of verbs and entities in knowledge using the proposed methodology, sentiments division has been matured. Great change occur in sentiments in case of *blood pressure* and *parkinson*. Positive sentiment increase from 19% to 75% and neutral sentiments changes from 69% to 8%for *blood pressure*. So the proposed system has analyzed that knowledge enhancer also effects sentiments division from tweets more precisely. As tweets are more precisely classified after knowledge enhancer and synonym binder so, later division of sentiments is more accurate.

V. CONCLUSION

In this research work, we have demonstrated a system to extract knowledge from tweets and then classify tweets based

on the semantics of knowledge contained in them. For avoiding information loss, knowledge enhancer is applied that enhances the knowledge extraction process from the collected tweets. The maturity of knowledge gained using knowledge enhancer module has helped to filter tweet more precisely avoiding information loss. We have also measured missing information during specific keyword-based search and then proposed a method to collect more precise information about specific topic or domain. Sentiment analysis shows people attitude towards different topics. This data can also help to generate richer user profile and generate valuable recommendations. In future we are planing to integrate the proposed system with personalized profile management, sentiment analysis, and recommender system.

ACKNOWLEDGEMENT

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2012-(H0301-12-2001)).

This work was supported by a grant from the NIPA(National IT Industry Promotion Agency) in 2013. (Global IT Talents Program).

REFERENCES

[1] "Alchemy api," (Last visited in March 2012). [Online]. Available: www.alchemyapi.com

[2] "American diabetes assosiation," (Last visited in october 2012). [Online]. Available: http://www.diabetes.org/diabetes-basics/common-terms/

[3] "Archivist," (Last visited in May 2012). [Online]. Available: http://archivist.visitmix.com/

[4] "Glossary of diabetes," (Last visited in october 2012). [Online]. Available: http://en.wikipedia.org/wiki/Glossary_of_diabetes

[5] "Grabeeter," (Last visited in November 2012). [Online]. Available: http://grabeeter.tugraz.at/

[6] "Healthcare tweet chats," (Last visited in October 2012). [Online]. Available: http://www.symplur.com/healthcare-hashtags/tweet-chats/

[7] "Java api for wordnet searching (jaws)," (Last visited in March 2013). [Online]. Available: http://lyle.smu.edu/~tspell/jaws/

[8] "The twitter hash tag: What is it and how do you use it?" (Last visited in January 2013). [Online]. Available: http://www.techforluddites.com/2009/02/the-twitter-hash-tag-what-is-it-and-how-do-you-use-it.html

[9] F. Abel, Q. Gao, G. Houben, and K. Tao, "Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web," in *Proceedings of ACM WebSci '11, 3rd International Conference on Web Science*. ACM, 2011.

[10] ——, "Analyzing user modeling on twitter for personalized news recommendations," *User Modeling, Adaption and Personalization*, pp. 1–12, 2011.

[11] F. Abel, Q. Gao, G.-J. Houben, and K. Tao., "Semantic enrichment of twitter posts for user profile construction on the social web," *The Semanic Web: Research and Applications*, pp. 375–389, 2011.

[12] S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using wikipedia," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 787–788.

[13] R. Batool, W. Khan, M. Hussain, J. Maqbool, M. Afzal, and S. Lee, "Towards personalized health profiling in social network," in *Proceedings of the 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM)*, 2012.

[14] V. Beal, "Twitter dictionary: A guide to understanding twitter lingo," (Last visited in October 2012). [Online]. Available: thttp://www.webopedia.com/quick_ref/Twitter_Dictionary_Guide.asp

[15] I. Celik, F. Abel, and G. Houben, "Learning semantic relationships between entities in twitter," *Web Engineering*, pp. 167–181, 2011.

[16] J. Dunn, "100 ways to use twitter in education, by degree of difficulty," (Last visited in October 2012). [Online]. Available: http://edudemic.com/2012/04/100-ways-to-use-twitter-in-education-by-degree-of-difficulty/

[17] N. Ellison *et al.*, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[18] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007, pp. 219–222.

[19] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 919–928.

[20] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American society for information science and technology*, vol. 60, no. 11, pp. 2169–2188, 2009.

[21] G. Kilpatrick, "The definitive list of twitter tools," (Last visited in November 2012). [Online]. Available: http://twittertoolsbook.com/the-definitive-list-of-twitter-tools/

[22] M. Naaman, J. Boase, and C. Lai, "Is it really about me?: message content in social awareness streams," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010, pp. 189–192.

[23] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003, pp. 70–77.

[24] I. News, "Social networks," (Last visited in October 2012). [Online]. Available: https://itunews.itu.int/En/512-Social-networks.note.aspx

[25] B. Rousseau, "Twitter statistics 2012 [infographic]," (Last visited in May 2012). [Online]. Available: http://gizmaestro.com/25/02/2012/social-media/twitter-statistics-2012-infographic/

[26] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2009, pp. 42–51.

[27] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*. ACM, 2010, pp. 841–842.

[28] M. Stankovic, M. Rowe, and P. Laublet, "Mapping tweets to conference talks: A goldmine for semantics," in *Workshop on Social Data on the Web, Shanghai, China*, vol. 664, 2010.

[29] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003, pp. 427–434.

[30] S. Zelikovitz and H. Hirsh, "Improving short text classification using unlabeled background knowledge to assess document similarity," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 1183–1190.