Microtext Processing

Richard Khoury
Department of Software Engineering
Lakehead University
Thunder Bay, Ontario
Canada
Richard.Khoury@lakeheadu.ca

Raphaël Khoury
Department of Computer Science and Software Engineering
Laval University,
Quebec City, Quebec
Canada
raphael.khoury.1@ulaval.ca

Abdelwahab Hamou-Lhadj
Department of Electrical and Computer Engineering
Concordia University
Montreal, Quebec
Canada
abdelw@ece.concordia.ca

Synonyms

Microtext: SMS, instant message, microblog, post, comment, status update, tweet.

Glossary

NLP: Natural Language Processing.

Definition

The term "microtext" was proposed recently by US Navy researchers (Dela Rosa and Ellen 2009) to describe a type of written text document that has three characteristics: (A) it is very short, typically one or two sentences, and possibly as little as a single word; (B) it is written in an informal manner and unedited for quality, and thus may use loose grammar, a conversational tone, vocabulary errors, and uncommon abbreviations and acronyms; and (C) it is semi-structured in the NLP sense, in that it includes some metadata such as a time stamp, an author, or the name of a field it was entered into. Microtexts have become omnipresent in today's world: they are notably found in online chat discussions, online forum posts, user comments posted on online material such as videos, pictures and news stories, Facebook newsfeeds and Tweeter updates, internet search queries, and phone text messaging (SMS).

The expression "microtext processing" refers to the branch of NLP that focuses on handling microtext. The processing tasks found within this branch overlap greatly with those in more traditional text processing areas, and include summarization, sentiment analysis, topic detection and classification, question-answering, and information extraction (Ellen 2011). However, new methodologies are being developed to accomplish these tasks by exploiting the unique text features of microtext highlighted above.

Introduction

The importance of microtext processing cannot be overstated. Billions of new microtexts are drafted every day. Furthermore, these microtexts are ripe with information, not only in their textual content but also in their associated metadata. This information is of value for NLP research, as well as for practical data mining application for social networking sites, web search engines, telecommunication companies, marketing firms, news sites, and many others.

The types and features of microtexts are directly dependent on the nature of the technological support that makes them possible. Consequently, the range of types and the specific features of each type will vary as new communication technologies are developed and become popular, and older technologies fade out. However, at this time, we can distinguish five main families of microtexts.

1. SMS texts: Messages sent through the Short Message Service (SMS) of cell phone networks are the most popular type of microtext in use today, with 8 trillion SMS sent worldwide in 2011, up from 6 trillion in 2010. The initial SMS support was limited to text messages of 160 characters to one recipient. However, most cell phone networks today support picture messages as well as text, while phones can handle messages of any length by breaking them up into shorter messages. It is also possible to send messages to multiple recipients, an option that has led both to business opportunities (such as joke-a-day texts) and to spam texting. This is the only type of microtext that does not require computer or internet access to be used. The metadata that can be collected from this microtext includes the author and recipient's phone numbers and geographic locations, and a timestamp.
2. Chat messages: These are public messages sent from one user to a group of people as part of a real-time group discussion. The users of the chat software usually do not know each other, although regular contributors can befriend each other over time. Moreover, the discussions are public, and can be read by anyone in the chat group as well as by anyone viewing the discussion silently (aka "lurking"). "Chat" initially referred specifically to Internet Relay Chat (IRC), and although that technology is still in use today it has been surpassed in popularity by other forms of chat, such as in-game chat rooms (such as the public chat channels in World of Warcraft) and website-based discussion forums. Typical metadata that can be obtained from chat messages include author, timestamp, and chat room information.
3. Instant Messages (IM): These are private messages sent from one user of an IM software to another as part of a real-time conversation. The fact that IM software today support group conversations has blurred somewhat the line between IM and chat, although some key differences remain, namely that group IM are between a private

group of friends while chat messages are between a public group that may not know each other at all. Typical metadata that can be obtained from IM include author, recipient, and timestamp.

4. Social Network posts: Although the idea of online communities is as old as the first networks that preceded the Internet, it is only in the early 2000s that these communities began to flourish, in part thanks to innovations such as personal profiles that allow users to list public interests and to compile lists of friends. Social networks allow a user to publicly post items of personal news, pictures, and videos, and to post comments on their friends' shared items. Social network posts are particularly rich in metadata, thanks to the users' profiles. In addition to the typical author, recipient, timestamp, and the exact item being commented on, one can collect social information about the author (their location, education level, income, etc.) and his/her relationship to the recipient (friend, colleague, friend-of-friend, etc.). The two most popular social networks today are Facebook and Twitter, and consequently we will focus on both of them in this work.
5. Web queries: These are the short messages used to search and retrieve information through web search engines. This type of microtext presents some unique features compared to the others. It is the only type of microtext that is not meant for a human recipient, and the only one that is always used for a single purpose, namely to obtain information from an automated system. Typical metadata that can be obtained from web queries include author, timestamp, and oftentimes either the user's entire or current session search history.

This chapter goes through a representative sample of microtext tasks. This first is topic detection and tracking, which includes multiple other tasks such as topic classification and clustering. Next, we will discuss spam detection, a very well-known task in text processing of immediate benefit to users. The third is text message normalization, a task that, by contrast to spam filtering, exists mainly in the realm of microtexts. And finally we will discuss sentiment analysis, a task that can only be done by NLP methods but which can be enriched by microtext metadata. Our discussion of these tasks will highlight both the similarities and differences between microtext processing and traditional text processing, and show how the metadata associated with microtexts can facilitate and enrich the processing tasks.

Key Points

Online text communications exhibit distinctive characteristics compared to regular documents: they tend to be very short, to be written in a loose and familiar style, and there is some amount of metadata associated with each text. These characteristics are sufficiently salient to warrant giving this class of "microtext" special attention. Indeed, it has been shown that, because of their unusual nature, microtexts cannot be effectively handled by traditional NLP algorithms. For example, the effectiveness of the Standford named entity recognizer algorithm falls from 90.8% to 45.8% when it is applied to a corpus of Tweets (Liu et al. 2001). In what follows, we illustrate how to develop NLP algorithms for microtexts that exploit, rather than are hindered by, their unique nature. We focus on a sample of four representative NLP tasks, namely topic detection and tracking, spam filtering, text message normalization, and sentiment analysis. We chose the first because it is a more complex task that requires solving other smaller challenges, the second because it is a common and well-known problem familiar to both researchers and laymen, the third because, by opposition, it is a problem that is considerably more present in microtext than other texts, and the fourth to illustrate how microtext features can be used to enrich an NLP task.

Historical Background

The appearance of microtext is a direct result of the development of telecommunication technology and of the Internet. Although microtexts were used as early as the 1980s, it was in the mid-1990s that their popularity exploded, as a result of the commercialization of the GMS mobile phone network with SMS support in 1993, and the release of numerous user-friendly IRC and IM software after 1995. The scientific community quickly took notice of these emerging text corpora, and began using them in research projects. We can initially differentiate two branches of research: on the linguistics side scientists performed research about microtext, while on the NLP side they performed research using microtext.

Some of the unique features of microtext, which are highlighted in Ellen's definition (Dela Rosa and Ellen 2009), were immediately obvious to linguists. In fact, as early as 1991, researchers studying what they called "interactive written discourse" on TELENET (a precursor of the Internet) had noticed that the messages exchanged there were shorter than in standard English and that they omitted pronouns, articles and copulas, used uncommon abbreviations, and featured incorrect capitalization (Ferrara et al. 1991). These researchers were among the first to note that these "e-messages", as they were called, should not be categorized as either written or spoken English but represented a new form of the language. Over the following decades, other linguists who studied chat messaging, SMS, and IM, echoed and expanded on these observations. New linguistic features, such as the use of phonetic substitutions ("u" for "you", "r" for "are") (Paolillo 1999) and the omission of punctuation (Baron and Ling 2007), were catalogued. Over time, the full social impact became appreciated: this was not a local trend between online friends but a socially prevalent (Paolillo 1999) and international (Baron and Ling 2007) phenomenon. However, referring to the definition in (Dela Rosa and Ellen 2009), these linguists overlooked the existence of metadata associated with the messages.

On the NLP side, the opposite approach was initially prevalent. Researchers sought to fit microtext corpora into the existing theoretical framework they were familiar with and use tried-and-true NLP methodologies. This typically meant preprocessing a corpus to make it more similar to a regular English text corpus. For example, when researchers in (Kolenda et al. 2001) wanted to implement a chat room topic detection method, they stripped all metadata from the chat messages and merged them all together into a single string, then arbitrarily split that string into what they called "pseudo-documents". This pre-processing allowed them to apply classical NLP methods to build document vectors for each pseudo-document and classify them. It wasn't before 2002 that NLP researchers began noting that chat conversations "differ […] in significant ways" from regular text (Wu et al. 2002). From there, they began to rediscover the linguistic features that linguists had been cataloguing for over a decade, and additionally noted that the surrounding metadata could be mined for information as well. So far as we can tell, it was four years later that the three key features of the texts – that they are short, written informally, and include metadata – were observed together for the first time, but specifically as important distinctive attributes of IM compared to regular text (Dong et al. 2006). It would be another three years before the suggestion was made that "microtext" was a separate class of text which encompassed IRC, IM, SMS, social network updates and more, and which was functionally defined by these three attributes (Dela Rosa and Ellen 2009). By that time, research about microtext was becoming common in the NLP branch of research as well, with new projects exploiting the unique features of microtext. Thus, in a sense, the NLP branch combined with and enriched the linguistic branch, and together they gave us this new field of "microtext processing".

Topic Detection and Tracking

The challenge of Topic Detection and Tracking (TDT) consists in monitoring a real-time source of information in order to detect the occurrence of a new event and to collect together all information related to this event. An event is defined, in this context, as a set of pieces of information that are related to the same topic and are highly concentrated in a period of time, and the occurrence of an event is an initial sudden spike in the number of pieces of information about this topic. Sources of microtexts have been found to be excellent resources to use for TDT, as they combine together messages generated in real-time by tens to hundreds of millions of users worldwide. Most sources of microtext could be used for that purpose: for example, Google Trends (http://www.google.com/trends/) gives a snapshot of the most popular current web search topics, and a spike in an unusual search topic is indicative of a new event related to that topic. However, microtext messages posted on social networks such as Facebook and Twitter are of particular interest here, since they are meant to be public messages spreading information about a given topic to a wide audience. Moreover, since the social network users commenting on on-going events are very interested and oftentimes actively involved in those events, and that they are commenting on-the-fly with no editorial oversight, news about events tends to spread both sooner and faster than it does in traditional media. Consequently, a growing body of research is dedicated to developing and studying new algorithms to detect social network posts related to the same event, and to pick out the occurrence of new individual events from the message stream as early as possible. The basic idea is to track the keywords used, and detect a sudden increase in the use of an unusual keyword by a large number of users.

The most popular NLP approach to detect the topic of text documents and to cluster documents related to the same topic is the bag-of-word approach. A word vector is built to represent the document, in which each entry of the vector represents the occurrence frequency of a word in that document. Similar documents can then be clustered together based on the distance between their word vectors, or a text document can be classified by computing the distance between its word vector and a class vector containing normal average word frequencies. This approach can then be refined in a number of ways, for example by using different metrics to compare the vectors, by assigning weights to the words, by using word ngrams or expressions, or by limiting the vector to important keywords, to name only a few. However, this approach was developed to deal with long text documents, and faces immediate problems when applied to microtexts: the distance between a heavily-populated class vector and the sparse word vector of a microtext will be unreliable, and the word vectors of two microtexts will have very little common vocabulary. A popular solution to this problem is to enrich the microtext by finding relevant additional keywords in an external text source. This has been done most successfully by submitting the microtext to a search engine and using the search results as new keywords. Other resources that have proved popular include WordNet, which can provide synonyms, antonyms, meronyms and holonyms of the words in the microtext, and Wikipedia, which provides a vocabulary classified in a large topic hierarchy.

Part of the challenge that arises when comparing and clustering social-network-generated microtexts is to deal with the many different ways that people can describe the same event, by using different words, expressions, and synonyms. This is a problem that is mitigated in longer text documents, such as news articles (another popular information source for TDT), where the length of the text insures a certain variety in the vocabulary and a good level of overlap between two articles describing the same news item. On the other hand however, the small number of words in microtexts exacerbates this problem. Consider for example these Facebook posts following the death of singer Amy Winehouse in 2011: “R.I.P. Amy Winehouse”, “We miss you Amy”, “I would do anything to bring Amy back”, “Amy Winehouse (1983-2011) RIP. Great Amy.”, “You and your music are in my heart Amy”. Clearly, aside from the given name “Amy”, there is very little vocabulary in common between these posts. Given the personal nature of social networks, it is natural that each person expresses his/her opinion in a different way. The fact that proper names of individuals, places, and organizations may be the only constant between posts has not gone unnoticed, and many systems have sought to place more weight on these proper nouns. Such systems can benefit from a Named Entity Recognition (NER) system. Given that social network posts will commonly refer to pop-culture knowledge and that the events the TDT system will need to detect will often be cultural ones, a good NER system in this context will be one that is enriched by popular culture names taken either from a general repository (such as Wikipedia) or from a domain-specific repository (such as IMDB).

The large set of problems encountered when trying to classify or cluster microtexts on the basis of their text content, combined with the fact that social network posts can be composed of non-textual content (such as videos, pictures, or external links), has led many authors to conclude that keyword-based approaches may not be a good strategy to use at all in this case (Takahashi *et al.* 2011). Consequently, researchers consider strategies to enrich the text with its associated metadata. One benefit of social-network microtext is the use of “hashtags” in the message to identify their topics with a single keyword. This was an innovation of Twitter, which encourages users to use them to clearly state the topic of their tweets, but using them became a trend on other sites; it is not uncommon for example to find hashtags in Facebook status updates, even though Facebook offers no hashtag support. Hashtags make it easy to detect and track topics, including both general news events (#breakingnews) and events related to specific topics (#amywinehouse). There have also been suggestions for TDT methods that exploit the linked nature of social networks. Linking is a foundational feature of social networks, and can be done by explicit mentions (e.g. naming another user in a post) and implicit mention (e.g. reposting or replying to another user’s post). Algorithms can look for anomalies, such as unusual increases in the number of mentions a post gets (Takahashi et al. 2011). Another alternative is to consider the overall shape of the post frequency distribution on a given topic. Indeed, the very nature of events is that they begin with a sudden sharp increase in mentions, prior to which they were seldom if ever discussed. This makes them fundamentally different from other topics such as trends and routines, which have much weaker peaks in mentions and are mentioned a lot more in-between peaks (Cvijikj and Michahelles 2011). The following figure illustrates this difference by comparing the frequency of posts mentioning “Amy Winehouse” around the day of her death, with the frequency of posts mentioning “Harry Potter” and “Happy Birthday” during the same period. As the figure shows, the event of Winehouse’s death causes a massive spike of about 5,000 posts per day, while the trendy topic of Harry Potter and the daily routine of wishing friends a happy birthday both show much more consistent behaviours, with day-to-day fluctuations of a few dozen posts to a few hundred posts at the most.

< microtext processing_fig1.png>

Figure 1. Frequency of posts mentioning "Amy Winehouse", "Harry Potter" and "Happy Birthday", taken from (Cvijikj and Michahelles 2011).

Spam Detection

The growth of modern communication technology has unfortunately been plagued by a matching growth in spamming, the mass-sending of unsolicited messages for commercial or malicious purposes. Spam has become ubiquitous in email communications, and as a result this platform has been the most studied. Most email spam detection methods work by abstracting the email as a bag-of-words or a vector of features extracted or constructed from the text, and then applying a classifier such as Naïve Bayes, support vector machine, or K-nearest neighbours. However, this strategy faces an immediate problem when dealing with microtexts. The small number of words in the microtext messages, along with the heavy use of non-standard vocabulary, makes the feature space larger and sparser than for longer and properly-written email messages. Consequently, this classification strategy is not quite reliable (Cormack et al. 2007), (Healy et al. 2005). One solution to consider is feature expansion: to create new text features in the messages, such as character frequencies (Healy et al. 2005) and orthogonal word bigrams (Cormack et al. 2007). The use of character frequencies makes sense for spam detection, since spammers often try to camouflage their messages from spam detection software by substituting one letter (e.g. V1AGRA) or by inserting punctuation (e.g. V.I.A.G.R.A.) (Healy et al. 2005). Orthogonal word bigrams are pairs of neighbouring but not adjacent words, such as "the fox" in "the quick brown fox". Listing all such orthogonal bigrams within a maximum neighbourhood distance does increase the number of features exponentially; but given how short the original text is even this larger enriched version remains well within acceptable time and space complexity bounds (Cormack et al. 2007). It is also possible to enrich the feature set using statistical text features, such as the proportion of uppercase to lowercase letters and the proportion of punctuation (Healy et al. 2005). In all cases, using some form of feature expansion has been found to improve the accuracy of microtext spam detection classifiers.

Feature expansion based on text is not the only option available. Microtext are rich in metadata that can be useful for filtering. Social network microtext is particularly rich in metadata, and spam filtering in that case can forego using the message text entirely. For example, some work has been done to quantify the "reputation" of users on social networking sites as the ratio of followers to friends that they have. It has been found that for a normal user this ratio is between 30% and 90%, while for spammer accounts it is an outlier value, either at 100% or below 20% (Wang 2010). Non-textual features of the messages can also be exploited, such as the frequency of sending duplicate messages, the frequency of messages that address a specific user by name, and the frequency of messages that include URL links, all of which are a lot higher than average for spammer accounts (Wang 2010).

Text Message Normalization

One of the fundamental characteristics of microtext is a highly relaxed spelling and a reliance on uncommon abbreviations and acronyms. This causes problems when we try to apply traditional NLP tools and techniques (such as Information extraction, automated summarization, or text-to-speech) that have been developed for conventional English text. It could be thought that a simple find-and-replace preprocessing on the microtext would solve that problem. However, the sheer diversity of spelling variations makes this solution impractical; for example, a sampling of Twitter (Petrovic et al. 2010) studied in (Liu et al. 2011) found over 4 million out-of-vocabulary words. Moreover, new spelling variations are created constantly, both voluntarily and accidentally.

The challenge of developing algorithms to correct the non-standard vocabulary found in microtexts is known as Text Message Normalization (TMN). The first step in tackling this challenge is to realize that, while the number of different spelling variations may be massive, they follow a small number of simple basic strategies (Liu et al. 2011):

1. Abbreviation: The user may delete letters (typically vowels) from the word. For example, in the Twitter corpus studied in (Liu et al. 2011), the word “together” was found sometimes rendered as “tgthr”.
2. Phonetic substitution: the user may substitute letters for other symbols that sound the same. This is typically done by using homophonic numbers, such as “2” for “to” in “2gether”.
3. Graphemic substitution: the user might substitute a letter for a symbol that looks the same. A common example is switching a letter “o” for the number “0”, such as in “t0gether”.
4. Stylistic variation: The user misspells the word to make it look more like its phonetic pronunciation (or sometimes specifically the user’s pronunciation). The Twitter corpus of (Petrovic et al. 2010) had several examples of this, such as using “togeda” or “togethor”.
5. Letter repetition: The user might repeat some letter for emphasis, for example typing “togtherrr”.
6. Typographic error: The user may have meant to type the word correctly, but made an honest mistake, such as swapping letters in “togehter". This is by far the most unpredictable strategy, as it follows no rules whatsoever. They are however also the most well-studied strategy, as typos are not unique to microtext but present in all typed text. As a result, there are a variety of common algorithms and off-the-shelf tools that can deal with them.

Naturally, these strategies are not mutually exclusive, and in fact they are frequently combined together. For example, the word “2gthr” is a result of abbreviation and phonetic substitution, while “toqethaa” comes from stylistic variation and letter repetition.

Several methods have been proposed to address the problem of TMN. They can be roughly categorized into three classes, corresponding to the three different metaphors that can be used to conceptualize the problem (Korbus et al. 2008). The first metaphor is the “noisy channel”, which models the non-standard tokens as noisy versions of the words they correspond to. In this metaphor, TMN can be seen as a special case of the spell-checking problem, and researchers can leverage the considerable amount of work that has already been published on that topic. The second metaphor is the “foreign language”, that instead considers the non-standard words of microtexts as the vocabulary of a foreign language that must be translated into English. As with the first metaphor, this second one allows researchers to benefit from the large set of existing methodologies developed for translation between natural languages for the purpose of TMN. The last metaphor is the “speech”, which stems from the observation that certain characteristics of microtext, such as the absence of reliable word separators and the tendency of spelling to mirror pronunciation, make it more akin to a spoken language than a written one. In this metaphor, one could exploit some existing speech recognition methods for TMN.

One last question that should be asked is, should microtexts be normalized into Standard English at all? One should consider the downsides as well as the benefits of this operation as well. The unconventional spellings of microtexts are not done only for cosmetic value, but they can also carry information (Baldwin and Chai 2011). Some users may prefer some of the transformation strategies to others, making them valuate tools for author identification. The spellings can also reflect the user’s emotional state; for example typing “yessssss” instead of “yes” can be used to express cheerfulness. TMN would discard all this information, and the resulting normalized text would thus not be semantically equivalent to the original.

Sentiment Analysis

As highlighted earlier, four out of the five types of microtexts serve to convey personal messages (the only exception being web queries, which constitute requests for information). Through these microtexts, people can share their opinions on various topics, from news reports to personal anecdotes. The challenge of sentiment analysis is to determine how users feel about the particular topics they are discussing. This can be greatly beneficial to a wide range of organizations – for example, to help companies understand how customers perceive a product or service, or to help political parties comprehend how voters feel about a candidate or issue.

Most sentiment analysis techniques rely on a combination of NLP methodologies and classification tools. From the NLP perspective, the detection of emotional keywords is of course a valid technique. It is also possible to use a Part-Of-Speech (POS) analysis to accurately model the emotional content of microtexts. N-gram models have also been used for this purpose; indeed, some authors report obtaining the best results with unigram models (Go et al. 2009) while others with bigram models (Pak and Paroubek 2010). Interestingly, in both cases the study was done using Tweets. This illustrates that an analysis based only on the text of microtexts can be unreliable. Consequently, researchers combine text analysis with classification tools to obtain better results. Indeed, the intensive use in microtext of emoticons – short ASCII strings that represent emotional faces, such as :) and :(for a smiling and a frowning face respectively – and of emotional response acronyms – such as LOL for Laughing Out Loud – gives features that can be immediately beneficial for traditional classifiers (Go et al. 2009).

The metadata of microtexts can of course be used to enrich sentiment analysis tools. In the case of Tweets, for example, it is possible to determine the sentiment of a message on the basis of features like retweets (users typically retransmit messages they strongly agree with) and hashtags (which can have emotional keywords, and group together tweets that often agree with each other). While these features alone are not sufficient for sentiment analysis, they have proven powerful additions to enrich traditional NLP methods (Barbosa and Feng 2010).

Key Applications

Microtext can be used in as large a range of applications as regular text, including for instance information extraction, automated summarization, and question answering (Ellen 2011). However, thanks to the fact that they are constantly being generated by users, microtext can also be used as a live stream of information in new types of social trend-monitoring applications. For example, microtexts (Twitter) can be used to measure multiple dimensions of political activity of users (Chen *et al.* 2012). A user's political engagement can be measured not only from the number of political messages but also from their features (the use of hashtags and retweets), while sentiment analysis techniques can be used to measure whether the messages are positive, negative or neutral. The authors were able to use this data to correctly predict the results of 8 out of 10 state elections in the 2012 “Super Tuesday” vote. Similarly, the authors of (Ritterman *et al.* 2009) used public perception of the 2009 H1N1 epidemic to predict stock market trends. They found that considering both the current number of posts on the epidemic and the historical trend over the past days and week gives a good picture of public opinion, which in turn allows them to model closely the movement of the stock market.

Future Directions

Although microtexts have existed for decades, it is only in recent years that they have gained notoriety, in part because of social and technological developments such as the rise of social networking. Given the sheer volume of microtexts in existence, the speed with which new ones are generated, and their versatility, we can only anticipate that this trend will continue in the future. Moreover, the real wealth of information in microtexts is not only found in the text content alone but also in the metadata that enriches it. This metadata can shed new light on interpersonal relationships in social networks, conversation dynamics in chat rooms, IM, and SMS, and patterns of human curiosity in web queries. Consequently, we expect that new algorithms will come along that will innovate in the type of metadata that can be analysed and in the information that can be understood from it. These developments will make microtext processing gradually branch off from traditional NLP. And as microtext processing algorithms mature, we can expect that they will begin to influence other areas of research. For example, one of the challenges in the development of ubiquitous systems is to make systems that are capable of correctly understanding user commands (short statements) that are very dependent on the user's current real-world situation (metadata). Likewise, the development of user interfaces that are more responsive to the user's current needs, of better automated question-answering systems, and in fact most systems that deal with human-machine interactions could benefit from advances in microtext processing.

Cross-References

Combining Online Maps with Text Analysis (00328)
Data Mining (00056)
Document Topic Identification (00352)
Multi-Classifier System for Sentiment Analysis and Opinion Mining (00351)
Sentiment Analysis (00120)
Social Web Search (00261)
User Sentiment and Opinion Analysis (00192)

References.

Baldwin, T., and Chai, J. Y., "Beyond Normalization: Pragmatics of Word Form in Text Messages", 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8-13 November 2011.

Barbosa L, Feng, J, "Robust sentiment detection on twitter from biased and noisy data", Proceedings of the 23rd International Conference on Computational Linguistics, pp. 36–44, 2010.

Baron, N., and Ling, R., "Text messaging and IM: Linguistic comparison of American college data", Journal of Language and psychological studies, vol. 26, pp. 291-298, 2007.

Chen, L., Wang, W., Sheth, A. P., "Are Twitter Users Equal in Predicting Elections? A Study of User Groups in Predicting 2012 U.S. Republican Presidential Primaries", SocInfo 2012, LNCS 7710, Springer, pp 379-392, 2012.

Cormack, G. V., Gómez Hidalgo, J. M., and Puertas Sánz, E., "Spam filtering for short messages", Proceedings of the 16th ACM Conference on Information and Knowledge Management (ACM CIKM'07). Lisbon, Portugal, pp. 313-320, 2007.

Cvijikj, I. P., and Michahelles, F., "Monitoring Trends on Facebook", Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, Zurich, Switzerland, pp. 895-202, 12-14 Dec. 2011.

Dela Rosa, K., and Ellen, J., “Text classification methodologies applied to micro-text in military chat”, Proceedings of the International Conference on Machine Learning and Applications, pp. 710-714, 2009.

Dong, H., Hui, S.C., and He, Y., “Structural analysis of chat messages for topic detection”, Online Information Review, 30(5), 496-516, 2006.

Ellen, J., “All about microtext: A Working definition and a survey of current microtext Research within Artificial Intelligence and Natural Language Processing”, ICAART (1), pp. 329-336, 2011.

Ferrara, K., Brunner, H., and Whittemore, G., “Interactive written discourse as an emergent register”, Written communications, vol. 8, pp 8-34, 1991.

Go, A., Bhayani, R., and Huang, L., “Twitter sentiment classification using distant supervision”, Technical report, Stanford, 2009.

Healy, M., Delany, S., and Zamolotskikh, A., “An assessment of case-based reasoning for short text messages”, Proceedings of the 16th Irish Conference on Artificial Intelligence and Cognitive Science, N. Creaney (ed.), pp. 257-266, 2005.

Kolenda, T., Hansen, L.K., and Larsen, J, “Signal detection using ICA: Application to chat room topic spotting”, Proceedings of the Third International Conference on Independent Component Analysis and Blind Source Separation, pp. 540-545, 2001.

Liu, F., Weng, F., Wang, B., and Liu, Y., “Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 2, pages 71-76, 2011.

Pak A., and Paroubek, P., “Twitter as a corpus for sentiment analysis and opinion mining”, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta: European Language, 2010.

Paolillo, J.C., “The virtual speech community: social network and language variation on IRC”, Proceedings of the 32nd Annual Hawaii International Conference on System Sciences, 1999.

Petrovic, S., Osborne, M., and Lavrenko, V., “The Edinburgh Twitter Corpus”, Proceedings of the NAACL HLT Workshop on Computational Linguistics in a World of Social Media, pp. 25-26, 2010.

Ritterman, J., Osborne, M., and Klein, E., “Using prediction markets and Twitter to predict a swine flu pandemic”, 1st International Workshop on Mining Social Media - 13th Conference of the Spanish Association for Artificial Intelligence, 2009.

Takahashi, T., Tomioka, R., and Yamanishi, K., “Discovering Emerging Topics in Social Streams via Link Anomaly Detection”, 11th IEEE International Conference on Data Mining, Tokyo, Japan, pp. 1230-1235, 11-14 Dec. 2011.

Wang, A. H., "Don’t follow me - Spam Detection in Twitter", Proceedings of the International Conference on Security and Cryptography (SECRYPT 2010) , pp. 142-151, 2010.

Wu, T., Khan, F.M., Fisher, T.A., Shuler, L.A., and Pottenger, W.M., “Posting Act Tagging using Transformation-Based Learning” The Proceedings of the Workshop on Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining (ICDM'02), December 2002.