TEXT ANNOTATION USING TEXTUAL SEMANTIC SIMILARITY AND TERM-FREQUENCY (TWITTER).

Abstract.

Researchers on social-media understandably assert that the contributions social media has made on various sectors is massive. Business development managers today have directed a huge amount of effort in strategizing efficient collaboration with both customers and other organizations using social-media. Despite the visible impact social media has made, a lot of digitally shared information is yet to be revealed. Gradually twitter has become the main hub for many Information system researchers, because tweets can freely be accessible in real-time by any one.

Motivated by earlier studies where IS researchers addressed big-data analysis and management by employing content analysis techniques, this paper proposes a novel approach to perform unsupervised classification of the tweets into different labels. It introduces a unique algorithm that uses semantic similarity between texts, Term-frequency and a determinant threshold to perform content analysis. The goal of this approach is to extract relevant features from a tweet thus reducing dimension and preparing training datasets that would be used to build classifiers.

Key-words: semantic-similarity, lexicon, text-mining, twitter.

1 Introduction

Prior scientific social media research that has harvested constructive knowledge from platforms such as twitter has often approached the topic of semantic analysis by making use of a variety of resourceful lexicon libraries or corpuses. They often use these to attach a sentiment label or class to a tweet, and this has proven to be a decent method of un-supervised classification of tweets into positive and negative classes. However, the problem of opinion mining goes beyond positivity and negativity. Occasionally the extent to which a text is positive or negative might not be enough to draw a conclusion, sometimes the analysis might aim to discover presence of subjects like "how knowledgeable people are", "how directive or useful people are". Such subjects and many more others might quite not be depicted by measuring the emotional strength of a statement. Additionally, sometimes, the suitable lexicon to fit the research or study might be inexistent and therefore, a need to manually create one arises. Manually creating lexicons is a cumbersome process requiring a lot of resources. (Wang, et al., 2016) went through three lengthy processes including word generation, word rating and psychometric validation to construct a stress word count dictionary, in their paper "Studying US Weekly Trends in Work Stress and Emotion".

Manual annotation of tweets has also gained huge popularity in the recent years. During the Sandy hurricane crisis that hit large parts of the Caribbean, the Mid-Atlantic and Northeastern United States in October 2012. (Brynielsson, et al., 2014) manually annotated tweets with one of the labels Anger, Fear, Positivity and Other in order to learn an optimum classifier whose aim was to improve the authorities' effectiveness of alert and communication towards the population during crises.

This paper addresses the text-annotation problem with a thorough, exhaustive and systematic technique termed textual semantic similarity labelling. Given any subject or keyword term, TSL looks for two or more words semantically similar to the keyword, and then calculates their term frequency upon which a tweet is labelled. The approach is tested and evaluated using a corpus of approximately 29,000 disease-related tweets upon which analysis was conducted to discover whether patients share remedies across the twitter platform.

2 Related Work

2.1 Linguistic Inquiry Word Count (LIWC)

(Chung and Pennebaker, 2012) breakdown LIWC, a word counting software program that references a dictionary of grammatical, psychological, and content word categories developed in 2007 by (Pennebaker, et al., 2001). Today, it's a widely used computational method for converting text into quantitative data or psychological constructs as phrased by some researchers. It is user-friendly, supports multiple languages, and has been applied to various fields including literature, personality, political science, etc. The logic behind LIWC is simply counting words and calculating word frequencies. Word frequency is calculated against a word count dictionary in terms of percentages, by using the following formula:

$$\text{LIWC word frequency} = \frac{wordcountsagainstadictionary}{totalwordcountinatext} * 100\%$$

(A dictionary needs to be defined beforehand, it refers to a collection of words and word stems (sometimes even phrases) that reflect or measure linguistic features or psychological constructs of research interest. For example, the LIWC "affect" dictionary comprises of 935 words and stems).

LIWC2015, the currently available version of the software, provides up to 100,000 text files containing over 250 million words. For a comprehensive list of the dictionaries, see the table "Comparing LIW2015 and LIWC2007" (Pennebaker Conglomerates, Inc., 2015). Among all its built-in dictionaries, the most

relevant to organizational health research are 19 dictionaries including social processes (e.g. family, friends, and humans), affective processes (e.g. positive emotion, negative emotion, anxiety, anger, and sadness), biological processes (e.g. body, health, sexual, and ingestion), and personal concerns (e.g. work, achievement, leisure, home, money, religion, and death).

Example of text analysis with LIWC

LIWC calculates the percentage of words in a text that matches a dictionary word, out of the total number of words in the text. For example,

1. In a single speech of 2000 words, that contained 150 pronouns and 84 positive emotion words, LIWC converts these numbers into percentages, 7.5% pronouns and 4.2% positive emotional words. (Pennebaker, 2015)

2. In a five-word text "I enjoyed my work today", the output by LIWC is 20% for the positive emotion dictionary (i.e. one positive emotion word "enjoyed" divided by five total words in the text, and multiplying by 100%), 20 (per cent) for the work dictionary (i.e. one work word "work"), and 0 (per cent) for the negative emotion and leisure dictionaries.

2.2 Language Assessment by Mechanical Turk (LabMT Sentiment Words)

During a study that aimed to characterize the content of tweets as well as the sentiments (average happiness) of US cancer patients. (Crannell, et al., 2016) adopted LabMT to aid an experiment that investigated the relative sentiment (happiness) as expressed in cancer patient tweets. They split a corpus of 186,406 tweets into different tweet-sets with each representing a specific cancer diagnosis, and further extracted individuals from the cancer-specific tweet-sets to form what would appear as cancer-patient tweet-sets. After cleaning up the datasets, they performed hedonometric analysis using the LabMT word list (Rinker, 2012).

LabMT is a word happiness list of the most frequently occurring 10,222 English words (Plos, 2011) compiled through frequency distributions from Google Books, the New York Times (1987-2007), music lyrics (1960-2007), and Twitter. To estimate the numerical average happiness (h_avg) of each word, words were scored on a 1-9 "happiness scale" using the popular online survey service Amazon Mechanical Turk. The happiest word is "laughter" (h_avg = 8.50), and the saddest word is "terrorist" (h_avg = 1.30).

The hedonometric analysis computed the average happiness value of a cancer-patient tweet-set by tallying the appearance of LabMT words found in the tweet-set. The average happiness value for a tweet-set is thus a weighted arithmetic mean of each word's frequency and the word's corresponding average happiness score. To increase the emotional signal, neutral "stop words" ($4 \leq h\_avg \leq 6$) are removed from the analysis.

2.3 MPQA (Multi Perspective Question Answering) lexicon

MPQA was constructed through human annotations of a corpus of 10,657 sentences in 535 documents that contain English-based news from 187 different sources in a variety of countries, dated from June 2001 to May 2002. The annotated lexicon represents opinions and other private states, such as beliefs, emotions, sentiments, speculations, etc. The subjective and objective expressions were also annotated with values of intensity and polarity to indicate the degree of subjectivity, and a negative, positive or neutral sentiment. The MPQA corpus contains a total of 8,221 words, including 3,250 adjectives, 329 adverbs, 1,146 any-position words, 2167 nouns, and 1,322 verbs. As for the sentiment polarity, among all 8,221 words, 4912 are negatives, 570 are neutrals, 2,718 are positives and 21 can be both negative and positive. In terms of strength of subjectivity, among all words, 5569 are strongly subjective words and other 2,652 are weakly subjective words.

(Ji, et al., 2016) is one of the many researchers that have utilized the MPQA opinion corpus described above to automatically assign labels or annotations to tweets. Their research used the subjectivity lexicon (sub-lexicon within MPQA), where a tweet was labelled Personal based on the assumption that if the number of strongly subjective clues and weakly subjective clues in the tweet is beyond a certain threshold (e.g. two strongly subjective clues and one weakly subjective clue), it can be regarded as a Personal tweet, otherwise it was labelled as a News tweet. They performed a verification step where tweets were examined again, if tweet contained at least three strongly subjective clues and at least three weakly subjective clues, it was labelled as a Personal tweet, otherwise, it was a New tweet.

For example, the tweet "Since when does a commercial aircraft accident become a matter of National Security Interests? #diegogarcia#mh370" is labelled Personal and "#UPDATE Cyanide levels 350x standard limits detected in water close to the site of explosions in China's Tianjin http://u.afp.com/Z5ab" is labelled News.

Other researchers have laboured through this task with more vigorous approaches like manually creating lexicons applicable to their research fields. (Nasukawa and Yi, 2003) demonstrate a procedure where they manually capture favourability to perform sentiment analysis. They categorize POS tags (nouns (NN), verbs (VB), adverbs (RB) and adjectives (JJ)) into positive and negative classes.

2.4 Emoticon-based annotation.

(Pak and Paroubek, 2010) queried tweets for two types of emoticons, Positive and Negative, in-order to attach sentiment to posts on 44 news paper's twitter accounts. Some of the emoticons they considered,

Happy emoticons: ":-) ", ":)", "=)", ":D" etc

Sad emoticons: ":-(", ":(", "=(", ";(" etc

Lexicon libraries or opinion corpuses are very powerful when correctly used to extract sentiment from textual data, they're a decent method of un-supervised classification of tweets into positive and negative classes. However, the problem of opinion mining goes beyond positivity and negativity. Occasionally the extent to which a text is positive or negative might not be enough to draw a conclusion, sometimes the analysis might aim to discover presence of subjects like "how knowledgeable people are", "how directive or useful people are". Such subjects and many more others might quite not be depicted by measuring the emotional strength of a statement.

Additionally, manually creating lexicons to fit the study or research questions if no dictionary exists is lengthy and often requires more resources, (Wang, et al., 2016) go through three lengthy processes including word generation, word rating and psychometric validation to construct a stress word count dictionary. Its' upon these arguments that SNADC adopts a procedure of measuring textual similarity to label tweets. Given the subject advise, TSL looks for two or more words semantically similar to the keyword advise, and then calculates their term frequency which is used to label a tweet as advise or not_advise.

3 Textual Semantic Similarity Labeling - TSL (Un-supervised classification).

3.1 Choice of class labels.

A class label is simply a name that identifies or represent a collection of related objects. Machine learning classification tasks aim to map an object or objects to a class or classes respectively. The experiments in this paper aimed to measure the extent to which a posted tweet provided advise related information i.e. two

keywords were used, Advise (the verb) and Advice (the noun). Four different class labels were identified, Purely_advise, Mostly_advise, Moderately_advise and Not_advise.

3.2 Using NLTK's WordNet (Wu-Palmer Similarity of synsets).

Borrowing (Zhai and Massung, 2016) understanding of Term clustering, A group of semantically similar terms are extracted from a tweet in the algorithm description that follows. Using the verb Advise, and the noun Advice as keywords, TSL traverses a group of tokens or individual words of a processed tweet, calculating the semantic similarity between the keywords and the tokens. That is, the semantic similarity between the verb "advise" and synonyms of each verb among the tokens was computed and similarly, that of the noun "advice" and all the nouns synonyms is computed. Synonyms are searched within synsets provided by NLTK's WordNet (a lexical database that groups English words into sets of synonyms called Synsets). For example, given a list of tokens S, **['diabetes', 'device', 'recommend']**, let A, B and C be computed similarities for the respective tokens in list S,

		Recommend		***Diabetes***	***Device***
synonyms	→	*recommend*	*commend*	*Diabetes*	*Device*
keywords	***Advise***	***0.22***	*0.16*	_	_
	Advice	_	_	***0.21***	***0.15***

The Synset of recommend, contains a set of its synonyms which are verbs 'recommend' and 'commend', and diabetes synset only has a noun 'diabetes' and device has a synset with just a noun 'device'.
SIM(A,B) = Wu-palmer similarity_measure between A and B

SIM(advise, recommend) = 0.22,
SIM(advise, commend) = 0.16
SIM(advice, recommend) = '_' incompatible POS for parameters, advice is a noun and recommend is a verb.
SIM(advice, diabetes) = 0.21

0.22 is greater than 0.16, therefore 0.22 is retained, i.e. The largest similarity measure among all computed similarities between key word and all synonyms is retained

A is 0.02, B, 0.2 and C,0.4 The output would be similarity_vector ***[0.22, 0.21, 0.15].***
(See Appendix A.1 for the python implementation of the Wu-palmer similarity function)

3.3 Feature extraction for training a classifier.

One-step threshold criteria is used to extract a smaller group of values that would represent the tweet when training a classifier. A tweet qualified to be annotated if (1) its similarity vector had two or more values. The threshold of two implied that tweets whose resulting token representatives were less than two words would get dropped. Finally, the algorithm extracted the largest three values to represent the tweet in the training set. Tweets that qualified but however had exactly two tokens only, would have missing values, the algorithm assigned a default value of 0 to make a vector of 3 values (Witten, et al., 2016).

Example if similarity_vector = [0.31, 0.02, 0.44, 0.00, 0.2], extracted feature set is [0.31, 0.44, 0.2], implying three most semantically similar words to the keywords.

3.4 Document Clustering using Term Frequency.

A text-summarization technique called TF-IDF is adopted to cluster the documents, TF-IDF is a weighting scheme widely used to measure the relative importance of a term/word in a large piece of text. (Alsaedi, et al., 2016) understand that TF-IDF approach requires knowing the frequency of a term in a document (TF) as well as the number of documents in which a term occurred at least once (DF). The semantic text automatic labeler evaluated the similarity vector of the entire group of tokens that represented the tweet independently instead of basing it on the whole corpus because the size of the corpus had diminished seriously at this step. Therefore, only the Term-frequency was used, the labeler searched for similarity measures that were either 0.2 or larger within the vector,

$$\mathbf{TF} = \frac{No\ of\ words\ with\ a\ similarity\ \ greater\ than\ 0.2}{Total\ No\ of\ terms\ in\ a\ similarity\ vector}$$

Finally, Tweets that had a Term-frequency weight less than 0.25 were labeled as not_advise, then those greater or equal to 0.25 but less than 0.5 were labeled moderately_advise, then mostly_advise for weights greater or equal to 0.5 but less than 0.75 and then purely advise for those with weight equal or greater than 0.75. The resulting training set is ready for supervised learning.

Continuing with the example above whose similarity_vector was [0.02, 0.2, 0.4]
TF = $\frac{2}{3}$ = 0.67, *therefore the tweet would be assigned* ***mostly_advise***.

*NB: Some tweets had several words whose similarity measure were greater than 0.2, i.e. e.g. for a similarity_vector = [**0.2**, **0.22**, 0.12, **0.35**, 0.01, **0.5**], The calculated* ***TF*** = $\frac{4}{6}$ *(0.67),*
*and extracted features for training set are the three largest i.e. [**0.22, 0.35, 0.5**].*

4 Experiment results and evaluation

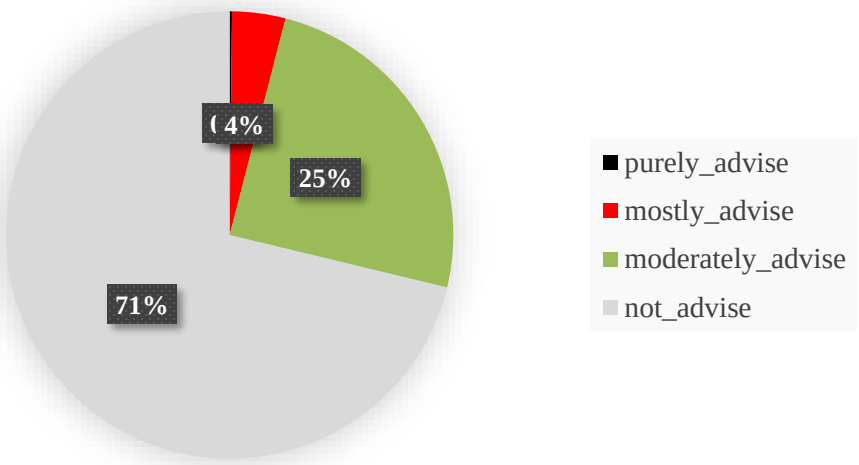

Figure 1: Chart showing the distribution of class labels across the diabetes dataset.

The statistics revealed by the chart show that the advice-related content found within the corpus of the disease related tweets is massively outnumbered by the content irrelevant to advice. Nonetheless, a sizeable number of tweets were discovered to contain some moderately advise related content. Despite the efficient communication and collaboration strategy social media has given birth to, much of the talk on social-media platforms is dominated by irrelevant issues about people's personal lives like lifestyle gossip, entertainment-industry scandals and links to videos and blog.

Tweet	Label
All #diabetics should be very aware, Of the importance of proper foot care https:\/\/t.co\/hHNsQgT09Z https:\/\/t.co\/SjVhRfdOQm #diabetes #poetry	purely_advise
Who says you can't look cool with #diabetes? \ud83d\ude0e\n\n#FridayFeeling (\ud83d\udcf8: @TDWsport)	not_advise
Your new improved EIO SmartCard is now on sale ! Available from https:\/\/t.co\/694ETPaplp #diabetes #epilepsy\u2026	purely_advise
Check this out Reverse Diabetes Naturally https:\/\/t.co\/y9pCDutXfh #diabetes	mostly_advise
RT @AmDiabetesAssn: The Senate released their #healthcare bill limiting access to care for ppl w\/ #diabetes. Tell your senators vote NO\u2026	not_advise
#Diabetes - what it is, how to prevent it, and how to manage it. https:\/\/t.co\/4OVjFJl9iI #diabetes #health #medicine #podcast	mostly_advise
RT @MSDintheUK: MSD is committed to #diabetes and proud to launch the social media campaign #T2DFirstThingsFirst. Follow us to learn more.\u2026	moderately_advise
Shocking study results - Reducing Sugar in Sodas Would Greatly Reduce Obesity and #Diabetes. https:\/\/t.co\/GnlU6mSkdM	moderately_advise

Table 1: Some of the results of annotated tweets using the textual semantic similarity algorithm

5 Conclusions.

Multiple re-usable frameworks that mine sentiment facts from big-data have been adopted in several case studies. Even though they have performed outstandingly and have been highly regarded in the past, we learn that the subject of opinion mining extends beyond classifying a tweet into positive and or negative classes. The techniques proposed in this paper provide a more robust approach that thoroughly assesses the opinion in a tweet by individually assessing each word in a tweet. This methodology adopted can scale to as many text-analytics applications as possible, because it's not limited by a dictionary or lexicon of words to guide the analysis, but rather simply follows a word by word analysis. Bench-marking of the tweet using text-summarization techniques such as TF-IDF and clustering ensured a thorough and efficient un-supervised classification procedure validating TSL as a text-annotation framework recommendable for usage in future text-mining applications.

References

Achrekar, H., Gandhe, A., Lazarus, R., Ssu-Hsin Yu and Liu, B. (2011). Predicting Flu Trends using Twitter data. *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*.

Alsaedi, N., Burnap, P. and Rana, O. (2016). Automatic Summarization of Real World Events Using Twitter. *International AAAI Conference on Web and Social Media*.

Antheunis, M., Tates, K. and Nieboer, T. (2013). Patients' and health professionals' use of social media in health care: Motives, barriers and expectations. *Patient Education and Counseling* 92(3), pp. 426-431.

Brynielsson, J., Johansson, F., Jonsson, C. and Westling, A. (2014). Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. *Security Informatics* 3(1).

Burnap, P., Gibson, R., Sloan, L., Southern, R. and Williams, M. (2015). 140 Characters to Victory? Using Twitter to Predict the UK 2015 General Election. *SSRN Electronic Journal*.

Chaffey, D. (2017). Global Social Media Statistics Summary URL: http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (visited on 19/07/2017).

Chung, C.K and Pennebaker, J. W. (2012). Linguistic Inquiry and Word Count (LIWC). *Applied Natural Language Processing*, pp. 206-229.

Crannell, W., Clark, E., Jones, C., James, T. and Moore, J. (2016). A pattern-matched Twitter analysis of US cancer-patient sentiments. *Journal of Surgical Research* 206(2), pp. 536-542.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*.

Ji, X., Chun, S. and Geller, J. (2016). Knowledge-Based Tweet Classification for Disease Sentiment Monitoring. *Sentiment Analysis and Ontology Engineering*, pp. 425-454.

Kane, G., Alavi, M., Labianca, G. and Borgatti, S. (2012). What's different about social media networks? A framework and research agenda. *MIS Quarterly* 38(1), pp. 275-304.

Manning, C., Raghavan, P. and Schütze, H. (2009). The text classification problem. In: *Introduction to information retrieval*. 1st ed. Cambridge: Cambridge University Press.

Nasukawa, T. and Yi, J. (2003). Sentiment analysis. *Proceedings of the international conference on Knowledge capture - K-CAP '03*.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *In Proceedings of the Seventh Conference on International Language Resources and Evaluation* 10, pp. 1320-1326.

Pennebaker Conglomerates, Inc (2017). LIWC 2015: Commparing LIWC 2015 and LIWC 2007 | LIWC [Online]. Available at: https://liwc.wpengine.com/compare-dictionaries/ [Accessed: 11 July 2017].

Pennebaker, J., Fancis, M. and Booth, R. (2001). Linguistic Inquiry and Word Count: LIWC2007. *Mahway: Lawrence Erlbaum Associates* 71.

Pennebaker, J. (2017). LIWC 2015: How it Works | LIWC. URL: https://liwc.wpengine.com/how-it-works/ [Accessed: 5 September 2017].

Plos (2011). *language assessment by Mechanical Turk 1.0*. URL:http://journals.plos.org/plosone/article/file?id=info%3Adoi/10.1371/journal.pone.0026752.s001&type=supplementary_(visited on 14/07/2017)

Rinker, T. (2017). labMT. qdapDictionaries 1.0.5 URL: http://trinker.github.io/qdapDictionaries/labMT.html (visited on 30/06/2017).

Wang, W., Hernandez, I., Newman, D., He, J. and Bian, J. (2016). Twitter Analysis: Studying US Weekly Trends in Work Stress and Emotion. *Applied Psychology* 65(2), pp. 355-378.

Zhai, C. and Massung, S. (2016). Term Clustering. In: *Text data management and analysis. A practical introduction to information retrieval and Text Mining*. 1st ed. ACM Book, pp. 284-287.