CONF-IRM 2019 Proceedings

International Conference on Information Resources Management (CONF-IRM)

5-2019

TOPICAL EXPRESSIVITY IN SHORT TEXTS

Herman Wandabwa
Auckland University of Technology, herman.wandabwa@aut.ac.nz

M Asif Naeem
Auckland University of Technology, mnaeem@aut.ac.nz

Farhaan Mirza
Auckland University of Technology, New Zealand, farhaan.mirza@aut.ac.nz

Russel Pears
Auckland University of Technology, russel.pears@aut.ac.nz

Recommended Citation

Wandabwa, Herman; Naeem, M Asif; Mirza, Farhaan; and Pears, Russel, "TOPICAL EXPRESSIVITY IN SHORT TEXTS" (2019). *CONF-IRM 2019 Proceedings*. 19.
https://aisel.aisnet.org/confirm2019/19

TOPICAL EXPRESSIVITY IN SHORT TEXTS

Herman Wandabwa
Auckland University of Technology
New Zealand
herman.wandabwa@aut.ac.nz

Farhaan Mirza
Auckland University of Technology
New Zealand
farhaan.mirza@aut.ac.nz

M. Asif Naeem,
Auckland University of Technology
New Zealand
mnaeem@aut.ac.nz

Russel Pears,
Auckland University of Technology
New Zealand
russel.pears@aut.ac.nz

Abstract

With each passing minute, online data is growing exponentially. A bulk of such data is generated from short text social media platforms such as Twitter. Such platforms are fundamental in social media knowledge-based applications like recommender systems. Twitter, for example, provides rich real-time streaming information. Extracting knowledge from such short texts without automated support is not feasible due to Twitter's platform streaming nature. Therefore, an automated method for comprehending patterns in such text is a need for many knowledge systems. This paper provides solutions to generate topics from Twitter data. We present several techniques related to topical modelling to identify topics of interest in short texts. Topic modelling is inherently problematic in shorter texts with very sparse vocabulary in addition to the informal language used in their dissemination. Such findings are informative in knowledge extraction for social media-based recommender systems as well as in understanding tweeters over time.

Keywords
Topic modelling, short texts, Twitter data analysis, recommender systems

1. INTRODUCTION

Microblogging on social media platforms like Twitter has emerged as de facto near real-time communication channels at marginal costs. Microblogging, as opposed to communication media, manage the conversational content as an artefact in the online environment (Di Grazia J, 2013). Increasingly, users leverage these platforms not only for interaction but in sharing news. A recent survey reveals that over 50% of microblog users consume news on Online Social Networks (OSNs) such as Twitter (Baldwin-Philippi, 2015). Twitter's popularity is unquestionable as it has over 326 million monthly active users. On average, 6000 tweets are sent per second, which corresponds to a dissemination capacity of over 350,000 tweets per minute and 500 million a day[1]. Tweeters[2] share hyperlinks, locations, photos and videos which inherently present challenges in terms of mining, processing, and understanding such data.

The ability for social media users to connect and interact anywhere and anytime, allows researchers to observe human behaviour at a different level, i.e. user social norms as well as

[1] *http://www.internetlivestats.com/twitter-statistics/*
[2] *A person who uses Twitter to update countless people on things they are doing at any given time.*

content in their interactions. Therefore, its possible to mine human behavioural patterns compared to traditional media. Microblogs have distinct characteristics that differentiate them from conventional long text platforms. Rich user interaction via multi-modal connections, varied relation types between users and content make them unique. A plethora of user-generated content that is dynamic, real-time and massive is generated. As a result, the big data problem is real in social media mining, i.e. "*drowning in data, but thirsty for knowledge*". Citizen journalism in the dissemination of event related data is one factor that has led to Twitter's growth [2]. The study of social patterns in such platforms helps identify collective and individual social patterns (Newman, 2009) among different online demographics. For instance, trending topics have been used in the detection of breaking news, content recommendation as well as targeted advertisements. User-Generated Content (UGC) is thus pertinent in microblogs as its valuable in applications such as event detection (Sakaki, 2010).

In this paper, we present several methods to learn topical interests' expressivity in short texts. Since v*ocabulary sparsity* is high in short texts, approaches, e.g. bag-of-words approach that works well on longer texts does not perform to the expected level on shorter texts. The contributions we made in this research are below: -

- Experiment the bag-of-words approach on short texts in the context of unigrams, bigrams and trigrams.
- Experiment on whether Term Frequency Inverse Document Frequency (TF-IDF) in the context of unigrams, bigrams and trigrams helps in extracting semantically better topics in short texts.

The rest of the paper is organised as follows. Section 2 summarises related literature. Section 3 describes our approach. In Section 4, we present our experimental framework and the results we obtained. Section 5 concludes the paper along with discussing some future issues.

2. RELATED WORK

Data in textual formats form a bulk of all online data. Much of this content is generated from short text data platforms like Twitter. Understanding and extraction of knowledge from such texts refer to topic modelling which is an active research area. Topic modelling algorithms such Latent Dirichlet Allocation (LDA) (Blei, 2003) and PLSA (Hofmann, 2017) have been successful in the discovery of latent topics in large text corpora in long text documents. However, the same algorithms do not perform well when subjected to short and noisy datasets (Yang, 2014) (Zhao W. X., 2011).

On the hindsight, LDA remains to be one of the popular algorithms of choice in uncovering latent semantic structures in such. Zhao et al. (Zhao W. X., 2011) modelled topics from tweets by assuming that each tweet belonged to one topic which may not be overly true. A tweet such as "*Maori culture should be entrenched in the education curriculum in New Zealand"* depicts more than one topic, i.e. *culture* and *education*. Therefore, the assumption that the contextual representation of such a tweet is a single topic based (single idea concept) was not a viable hypothesis.

Topical key phrases extraction methodology for tweets was proposed by some researchers (Zhao W. X., 2011). Their methodology involved three main processes; generation and ranking of keywords and key phrases ranking based on the ranked keywords. In ranking keywords, Topical PageRank method (Liu, 2010) was modified by introducing a topic sensitivity related score. The topical key phrases were ranked via principled probabilistic phrase ranking. User interests in the approach were modelled via the retweeting patterns using this key phrase ranking methodology. Better topical key phrases were extracted from short texts using this approach.

In terms of finding topic-sensitive influential tweeters (Weng, 2010) proposed aggregation of all tweets disseminated by each tweeter over time in one document. LDA was applied to the document to extract the user's interests in the text. However, sparsity in word co-occurrence, as well as noise in the text, led to the generation of bad topics.

Understanding short texts with high vocabulary sparsity via external knowledge is another thought process that some researchers adopted. (Andrzejewski, 2009) incorporated external domain knowledge in the conventional LDA modelling process via Dirichlet Forest priors (DF-LDA). The idea was that knowledge could be expressed with two primitives on word pairs, i.e. *Must-Links* and *Cannot-Links*. This represented words that could co-occur and those that couldn't. Words like *"Auckland, museum"* or *"English, language"* may always co-occur in the topic modelling process. This co-occurrence knowledge was encoded as Must-Links in the Dirichlet Forest prior to augment the LDA process for extraction of coherent topics. Other researchers proposed a semi-supervised approach called Metamodel Enabled Latent Dirichlet Association (MELDA) where a metamodel was built from semantically relevant long texts to guide the LDA process in short texts (Wandabwa, 2018).

Therefore, topic modelling in short texts is a challenging task. Several variants of topic modelling algorithm have been proposed by several researchers above. External knowledge in the form of metadata, aggregation of short texts in one document as well as studies related to the social structure of short text platforms are some of the approaches that have been used in knowledge extraction in short texts. Our technique involved usage of TF-IDF based scoring algorithm in learning the relevance of words relative to topics for expressivity.

3. OUR APPROACH

Our modelling approach is built upon LDA. LDA is described as an unsupervised approach in textual data knowledge discovery. The method generates a summary of pre-set topics through a discrete probability distribution over words. It then infers per-document distribution over the generated topics. Each document is therefore interpreted as a mixture of various topics with the topic distribution assumed to have a sparse Dirichlet prior. This sparsity ensures that documents only cover a small set of topics and that topics can also be captured by a small set of words that reduce ambiguity in the generated set of topics (Blei, 2003). Formulation of a topic in LDA is based on term co-occurrence likelihood which is core in our approach as much as the dataset is short text based. Therefore, we make use of N-Gram modelling in the generation of vocabulary and TF-IDF in assigning importance to the generated vocabulary. This mitigates the issue of vocabulary sparsity which is prevalent in LDA.

4. METHODOLOGY

4.1. Bag-of-Words

In this approach, the text is represented as a multiset of words disregarding order and the grammatical structure but upholding multiplicity. Normally, it involves a vocabulary of known words and a measure of the presence of the known words. Each word count is considered a feature (Goldberg, 2017). The intuition in this approach is that documents with a high similarity of words are semantically close.

The complexity of a topic modelling algorithm depends on two factors: -

- The methodology used in choosing the vocabulary from known words. This is important in vocabulary with high sparsity.
- The scoring technique for words in the dictionary. This relates to the importance of word co-occurrence patterns in the dictionary.

In our approach, a simple method of scoring words in the vocabulary was applied. The presence and absence of a word in the vocabulary as compared to a single document (tweet) was assigned a boolean value of 1 or 0 respectively after vectorisation. New tweets that overlap with the vocabulary of known words are still encoded. For model consistency, words in new tweets that are not present in the vocabulary are simply discarded. Due to the presence of many tweets, vector representation of tweets increases exponentially.

4.2. N-Gram and TF-IDF Modelling

Patterns characterize sentence construction in textual data. Therefore, word order is pertinent. For some algorithms to learn word order sentences, probabilities in word ordering need to be computed. Therefore, an N-gram is a sequence of *N* words. A two-word sequence is thus a bi-gram and a three-word sequence, a tri-gram.

Computing the probability of a word *w* given a document word sequence history *x* can be naively computed from relative counts. In a sentence history such as *"Belgium should be the champions*" as *x*, a count of the number of times *x* is followed by *w*, i.e. "champions" divide by the number of times *x* is not followed by *w*.

In our case, the number of documents/tweets was high. We, therefore, used joint probabilities of sentences whereby sequences of words are computed based on the number of words in the sequence divided by counts of all possible sequences of the available words. To estimate the probability of a word w given the history x of sentence sequences, an approximation of x to w is computed by looking at the last few words and not the entire sequence in the history.

In the bi-gram model, for example, an approximation of the probability of a word given all previous words *n* is

$$P(W_n | W_1^{n-1})$$

To predict the next word in such a model, the approximation of the same is as below: -

$$P(W_n | W_1^{n-1}) \approx P(W_n | W_{n-1})$$

The tri-gram model in this instance is generalised to check for two previous words as compared to one previous word in the bi-gram model. Estimating the N-gram approximation to the conditional probability of the next word is computed as below: -

$$P(W_n | W_1^{n-1}) \approx P(W_n | W_{n-N+1}^{n-1})$$

Regarding TFIDF, the assumption is that each tweet *d* is treated as an individual document in the corpus. The document is viewed as a vector with one component (a word) in the dictionary, together with a weight for each word. In TF-IDF weighting scheme, a word *w* in document *d* is weighted as below: -

$$tf\_idf_{w,d} = tf_{(w,d)} \times idf_w$$

The computed term/word weight in a document is highest when (w) occurs many times within a small number of tweets (t). The discriminating power to those documents is highest. The weight is lower when the word occurs fewer times in a document or many documents. It is lowest when the term occurs in all documents.

4.6. EXPERIMENTAL SETUP

We present quantitative and qualitative results of our model in this section. Our experimental setup consisted of several processes to test the models described in Section 5.

4.7. Datasets and Settings

4.7.1. Dataset

Our interest is in the extraction of knowledge from tweets by looking at the topics they present. We collected FIFA World cup 2018 related tweets for a period of one month starting 15/06/2018 to 16/07/2018, dates coinciding with the competition via Twitters streaming API. The data provided a good scenario to test our approaches based on the diverse topics within the dataset. We collected 2,894,926 unique tweets for analysis. Each entry in the dataset represented a single tweet with its associated metadata, e.g. geo, mentions, hashtags etc.

4.7.2. Data Pre-processing, Parameter Settings and Model Tuning

Data pre-processing in our approach entailed cleaning up symbols, numbers, punctuations as well as lowercasing the textual bit of the tweets. Parameters in modelling were setup to optimise the modelling process as well as in the generation of better topics. LDA was used in the topic modelling process. However, the difference in its output was based on the generated dictionary which is the input of the model. The dictionary was generated as unigrams, bigrams and trigrams. We then applied LDA on these N-gram variances as well as with TF-IDF.

For unigrams, bigrams and trigrams, the minimum count was set to 3. It meant that all words and bigrams with a total count of less than three were ignored. A score threshold for forming phrases was set at 100. A phrase of words *x* followed by *y* is accepted as one if the value is higher than the threshold meaning the higher the value, the fewer the phrases. For LDA modelling, the (number of topics) was 10, with two *passes* in the dictionary. The same parameters were replicated with TF-IDF applied to the dictionary.

4.8. Results

4.8.1. Quantitative Evaluation of Topics

Some objective metrics determines the quality of topics in such models. We applied topic coherence as the topic's evaluation metric (Mimno, 2011). The choice of this metric was with relation to how interpretable the extracted topics were to humans. Higher coherence values meant better topic quality thus better topic discernment by humans. Results below show that the TF-IDF-based LDA model with trigrams consistently outperformed the other models as per the values in Figure 2 below. A higher value means better topic quality.

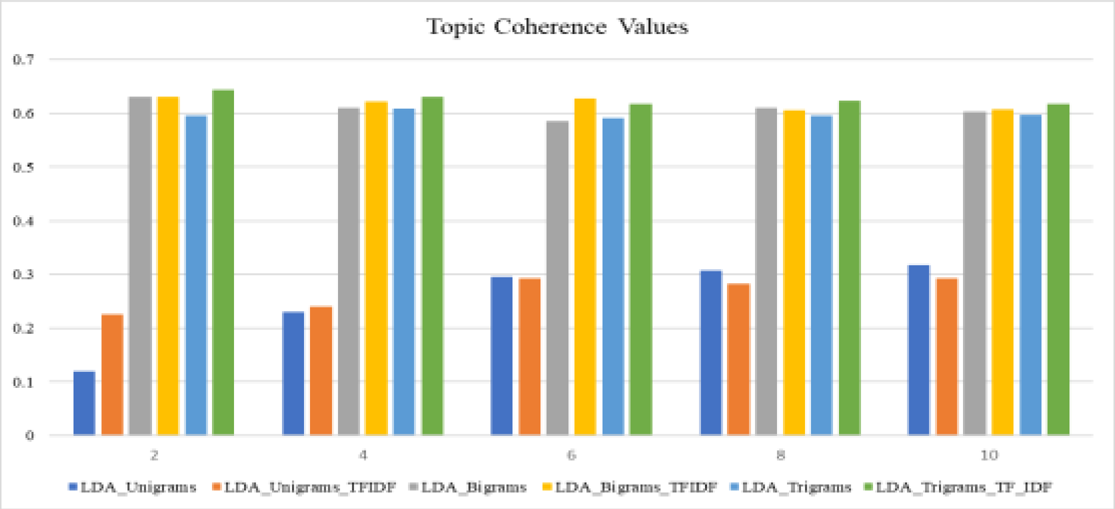

Figure 1: Plot showing the coherence values at different topic numbers.

4.8.2 Qualitative Evaluation of Topics

The main goal of the topic modelling is to generate topics that are interpretable to humans. Quantitative assessment of generated topics is not enough on its own. Therefore, human judges were chosen to discern the quality of generated results based on unseen documents. To do this, we randomly selected three judges with enough English articulacy to label every generated topic. It is worth noting that the judges are unrelated to the authors and thus not considered biased. We choose the top 20 words ranked by per-topic word distributions for each topic. A word was considered relevant to the topic if all judges agreed to it; otherwise was labelled as bad. The remaining ones were labelled as neutral, i.e. did not affect the expressivity of the topics. For transparency, topic model names were not given but just placeholders. Results from this evaluation are presented in Figure 3 inform of a Kappa Score. A topic model is deemed good if 60 percent of the assessments from the judges deemed it so. From the results in Figure 2, the TFIDF-based LDA model on tri-grams outperformed the rest of the models in 8 and ten topics. TFIDF-based LDA model on bi-grams also performed well on when subjected to ten topics. Its score was lesser when subjected to 8 topics. We chose 8 and ten topics based on the consistency in the previous result.

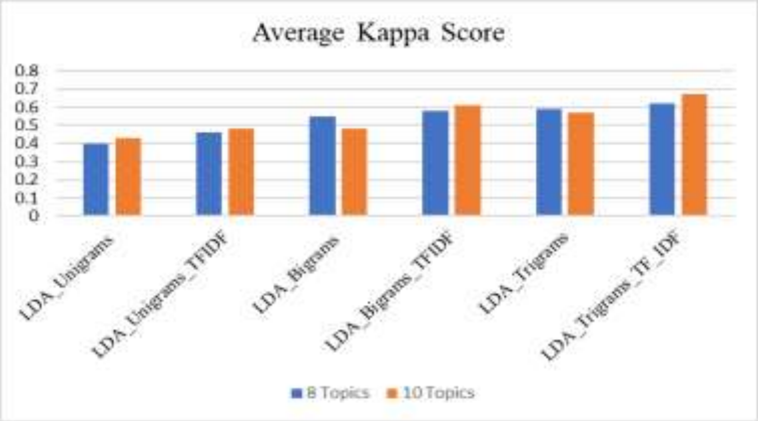

Figure 3: Average Kappa Score as evaluated by humans

Table 1 below presents a sample of topics and respective keywords within the topics when they are tested with an unseen sentence.

Unseen Document: *“Croatia game versus France should have had three red cards. Great final”*		
Model	**Score**	**Top 10 keywords**
LDA-Trigram- TFIDF	**0.8875**	*France, Croatia, worldcupfinal, fracro, final, world, frabel, England, Belgium, congratulations*
LDA-Bigrams	**0.8875**	*France,Croatia,worldcupfinal,fracro,Iceland,Belgium,frabel,argentina,Switzerland, argisl"*
LD-Unigrams-TFIDF	**0.7240**	*France, Croatia, Belgium, England, worldcupfinal, final, fracro, Mexico, world, croeng*
LDA-Trigrams	**0.6837**	*Croatia, France, goal, worldcupfinal, fracro, great, peru, argcro, best, final*
LDA-Unigrams	**0.6582**	*France, worldcupfinal, Croatia, korea, Sweden, south, fracro, Mexico, Belgium, germany*
LDA-Bigrams	**0.6418**	*France, Croatia, worldcupfinal, Belgium, fracro, frabel, beljpn, match, world, iran."*

Table 1: Sample topical relevance scores of topic models on an unseen document

5. Conclusion

The emergence of social media has led to an enormous volume of online data. Short texts such as tweets present challenges when it comes to extraction of topics. This is largely because of high sparsity in word co-occurrences. In this paper, we presented several LDA-based techniques for modelling topics in short texts. As much as LDA's performance has not been optimal in discerning short text topics, tweaks to the dataset generated more interpretable topics. In this study, a TFIDF-based LDA model on tri-grams performed consistently well in the extraction of the topics. Fine-tuning of the model input parameters via TF-IDF makes a big difference in the topic's generation process. As part of the future work, we are interested in exploring usage of word embeddings in modelling topics regardless of the linguistic structure.

References

Andrzejewski, D. Z. (2009). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *26th annual international conference on machine learning* (pp. 25-32). ACM.

Baldwin-Philippi, J. (2015). *Using technology, building democracy: Digital campaigning and the construction of citizenship.* Oxford University Press.

Blei, D. M. (2003). Latent dirichlet allocation. *Journal of machine Learning research*.

DiGrazia J, M. K. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *Plos One* .

Goldberg, Y. (2017). Neural network methods for natural language processing. . *Synthesis Lectures on Human Language Technologies,* , 1-309.

He, Y. W. (2018). Discovering canonical correlations between topical and topological information in document networks. . *IEEE Transactions on Knowledge and Data Engineering*, 460-473.

Hofmann, T. (2017). Probabilistic latent semantic indexing. *ACM SIGIR Forum* (pp. 211-218). ACM.

Liu, Z. H. (2010). Automatic keyphrase extraction via topic decomposition. *2010 conference on empirical methods in natural language processing* (pp. 366-376). Association for Computational Linguistics.

Newman, N. (2009). *The rise of social media and its impact on mainstream journalism.*

Sakaki, T. O. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. . *19th International conference on World Wide Web* (pp. 851-860). ACM .

Wandabwa, H. N. (2018). A Metamodel Enabled Approach for Discovery of Coherent Topics in Short Text Microblogs. . *IEEE Access*.

Wang, X. W. (2011). Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. *20th ACM International Conference on Informationand Knowledge Management* (pp. 1031–1040). ACM.

Weng, J. L. (2010). Twitterrank: finding topic-sensitive influential twitterers. *Third ACM international conference on Web search and data mining* (pp. 261-270). ACM.

Yang, S. H. (2014). Large-scale high-precision topic modeling on twitter. *20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1907-1916). ACM.

Zhao, W. X. (2011). Comparing twitter and traditional media using topic models. In. *European conference on information retrieval* (pp. 338-349). Berlin, Heidelberg.: Springer.

Zhao, W. X. (2011). Topical keyphrase extraction from twitter. *Proceedings of the 49th Annual Meeting of the Association for ComputationalLinguistics* (pp. 379-388). Association for Computational Linguistics.