

Common Topic Identification in Online Maltese News Portal Comments

Samuel Zammit^a, Fiona Sammut^b and David Suda^c

Department of Statistics & Operations Research, University of Malta, Msida, Malta

Keywords: Natural Language Processing, Word Embeddings, Word2Vec, FastText, Doc2Vec, *k*-means Clustering.

Abstract: This paper aims to identify common topics in a dataset of online news portal comments made between April 2008 and January 2017 on the Times of Malta website. By making use of the FastText algorithm, Word2Vec is used to obtain word embeddings for each unique word in the dataset. Furthermore, document vectors are also obtained for each comment, where again similar comments are assigned similar representations. The resulting word and document embeddings are also clustered using *k*-means clustering to identify common topic clusters. The results obtained indicate that the majority of comments follow a political theme related either to party politics, foreign politics, corruption, issues of an ideological nature, or other issues. Comments related to themes such as sports, arts and culture were not common, except around years with major events. Additionally, a number of topics were identified as being more prevalent during some time periods rather than others. These include the Maltese divorce referendum in 2011, the Maltese citizenship scheme in 2013, Russia's annexation of Crimea in 2014, Brexit in 2015 and corruption/Panama Papers in 2016.


1 INTRODUCTION


As the use of the internet and online social media increases, text data is becoming an ever more important source of data. Therefore, it is no surprise that Natural Language Processing (NLP) tools and techniques have seen a rapid increase in use in recent years. The applications of these techniques are varied in scope; they include sentiment analysis (Socher et al., 2013b), detection of political biases (Iyyer et al., 2014), and extracting relationships between words (Mikolov et al., 2013b). A number of research papers have been written on the subject of news content and online comment analysis using NLP. These include the analysis of newspaper articles (Costola et al., 2020), comments on online news portals (Zaidan and Callison-Burch, 2014) and identification of spam comments (Rădulescu et al., 2014).


Seminal models in NLP, such as *n*-grams and the Bag-of-Words model, suffer from the curse of dimensionality. Neural networks can overcome the dimensionality problem posed by standard techniques. Through the use of neural networks, the high-dimensional discrete vector representation attributed to each term is replaced by a lower-dimensional con-

tinuous vector representation, known as a word embedding (Bengio et al., 2003). A further advantage of these models is that unseen words can be modelled by using the surrounding words as context for the model. Extensions and improvements to the first word embedding neural network models include the now-famous Word2Vec (Mikolov et al., 2013a) and FastText (Bojanowski et al., 2017) models, which provide improved results at a lower computational cost. Methods for obtaining document embeddings, as opposed to embeddings for individual words, have also been developed (Le and Mikolov, 2014; Mouselimis, 2019). The obtained vector representations may be analysed using a variety of techniques, including clustering and similarity metrics.

The aim of this paper is that of common topic identification in Times of Malta (ToM) website comments made between April 2008 and January 2017. This is the first study of its kind on Maltese online commentary, and also unique in the approach it uses. These results may be used by news portals in order to determine which topics generate the most engagement among readers, and hence generate more clicks (and consequently more revenue). The structure of this paper is as follows. In Section 2, we discuss *n*-grams and skip-grams, which are fundamental concepts often used in NLP and which are needed within the word embedding context. In Sec-

^a  <https://orcid.org/0000-0002-3290-8684>

^b  <https://orcid.org/0000-0002-4605-9185>

^c  <https://orcid.org/0000-0003-0106-7947>

tion 3, we describe an approach that can be used to obtain vector representations, or embeddings, of words. In particular, we describe Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and FastText (Bojanowski et al., 2017). This idea is also extended to documents through the use of Doc2Vec (Mouselimis, 2019). Section 4 then details the application section of this paper. Firstly, the most common n -grams are extracted for the comments dataset. Consequently, an embedding is obtained for each unique word in the dataset, and these embeddings are then clustered using k -means clustering, where each cluster roughly corresponds to a topic of discussion. The same exercise is repeated to obtain document vectors, where each comment is treated as a separate document. Different subsets of the data are considered in order to identify trends in discussion topics over time. The results of the clustering are presented in this section. Lastly, Section 5 sums up a conclusion of the paper, including limitations encountered during the study, as well as detailing related and possible future research work on the area.

2 N-GRAMS AND SKIP-GRAMS

The concept of n -grams lies at the foundation of most NLP theory. An n -gram is defined as a sequence \mathbf{w}_1^n of n words

$$\mathbf{w}_1^n = (w_1, w_2, \dots, w_n)' \quad (1)$$

within a document d made up of $|d| \geq n$ words. For example, in the sentence ‘‘I took my dog to the park’’, the possible n -grams are as follows:

- Possible unigrams: {I, took, my, dog, to, the, park},
- Possible bigrams: {I took, took my, my dog, dog to, to the, the park},
- Possible trigrams: {I took my, took my dog, my dog to, dog to the, to the park},

and so on. The probability of the n -gram \mathbf{w}_1^n in equation (1) is given by

$$\mathbb{P}(\mathbf{w}_1^n) = \mathbb{P}(w_1)\mathbb{P}(w_2|w_1)\mathbb{P}(w_3|\mathbf{w}_1^2) \dots \mathbb{P}(w_n|\mathbf{w}_1^{n-1}). \quad (2)$$

While any n -grams that do not appear in the training text can be assumed to be very rare, their aggregate probability should be taken into account, and it is not correct to assume zero probabilities of occurrence (Brown et al., 1992). If an n -gram is assigned a probability of zero, any sequence of words containing this n -gram will be incorrectly assigned probability zero due to equation (2).

An approach which may be used to mitigate the zero probabilities encountered in using n -grams is the use of skip-grams. Skip-grams, also known as skip n -grams (Pickhardt et al., 2014), allow context to be ‘skipped’. The set of k -skip- n -grams for the N -sequence \mathbf{w}_1^N is defined by

$$\left\{ (w_{i_1}, w_{i_2}, \dots, w_{i_n}) \text{ s.t. } \sum_{j=1}^n (i_j - i_{j-1}) \leq k + 1 \right\}, \quad (3)$$

where $1 \leq i_1 < i_2 < \dots < i_n \leq N$. Note that $i_j - i_{j-1}$ in equation (3) is the difference between two subscripts corresponding to words that are adjacent in a k -skip n -gram. Therefore, if $i_j - i_{j-1} = k$, the words w_{i_j} and $w_{i_{j-1}}$ are k words apart in \mathbf{w}_1^N , i.e. there are $(k - 1)$ skips between the two words.

Using the same example sentence as for the n -grams example, we illustrate some skip-gram examples:

- Possible 1-skip-bigrams: {I took, I my, took my, took dog, my dog, my to, dog to, dog the, to the, to park, the park},
- Possible 2-skip-bigrams: {I took, I my, I dog, took my, took dog, took to, my dog, my to, my the, dog to, dog the, dog park, to the, to park, the park},
- Possible 1-skip-trigrams: {I took my, I took dog, I my dog, took my dog, took my to, took dog to, my dog to, my dog the, my to the, dog to the, dog to park, dog the park, to the park},

The probabilities of words and sequences of words may be worked out in the same way as with non-skip n -grams, except that the set of n -grams is much larger when allowing for skips in context. Skip-grams lead to an increase in coverage (Guthrie et al., 2006) and may be used to tackle the data sparsity problem, for example when the training text is limited in size (Allison et al., 2006). However, this increase in model size may also be seen as a potential disadvantage, since a model that is excessively large may cause computational issues (Pibiri and Venturini, 2019).

3 EMBEDDING VECTORS

In order to get around the curse of dimensionality problems caused by traditional language models, we consider neural language models. In particular, we convert each word into a vector based on the contexts in which this word appears, creating what is known as a distributed representation of the word, or word embedding. A neural language model goes through the following steps to process a word w_i :

- Each word w_i is fed into the neural network as a $1 \times h$ -dimensional one-hot encoded vector \mathbf{w}_i ;
- \mathbf{w}_i is multiplied by the $h \times q$ word embedding matrix $\mathbb{W}^{(H)}$ and transformed using an activation function $\sigma_1(\cdot)$ to obtain the hidden layer $\mathbf{a}_i^{(H)}$;
- $\mathbf{a}_i^{(H)}$ is multiplied by the $q \times h$ output weight matrix $\mathbb{W}^{(O)}$ and transformed using an activation function $\sigma_2(\cdot)$, typically the softmax function as in equation (5), to obtain the output layer.

In neural word embedding models, only $\mathbb{W}^{(H)}$ is used, while $\mathbb{W}^{(O)}$ is discarded after training. Then, the embedding \mathbf{v}_i for the word w_i is given by $\mathbf{v}_i = \mathbf{w}_i \mathbb{W}^{(H)}$. In the following subsections, two commonly-used neural language models are described. These are Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and FastText (Bojanowski et al., 2017), which were developed by research teams at Google and Facebook respectively.

3.1 Word2Vec

Word2Vec consists of two ‘opposing’ model architectures, namely the continuous bag-of-words (CBOW) model and the continuous skip-gram model (Mikolov et al., 2013a). A main advantage of Word2Vec over previous neural word embedding models (Bengio et al., 2003; Bengio and LeCun, 2007; Mikolov et al., 2010) is a reduction in computational complexity. This shall be seen through the use of negative sampling explained at the end of this subsection. We shall focus mainly on the continuous skip-gram model. The skip-gram loss function is given by the negative loglikelihood, namely

$$L = -\frac{1}{|V|} \sum_{i=1}^{|V|} \sum_{\substack{j=-c \\ j \neq 0}}^{j=c} \ln \mathbb{P}(w_{i+j}|w_i), \quad (4)$$

where $|V|$ is the vocabulary size, c is the context window size and $\mathbb{P}(w_{i+j}|w_i)$ in equation (4) is defined by the softmax function

$$\mathbb{P}(w_{i+j}|w_i) = \frac{\exp(\mathbf{v}_{i+j}^* T \mathbf{v}_i)}{\sum_{v=1}^{|V|} \exp(\mathbf{v}_v^* T \mathbf{v}_i)}. \quad (5)$$

A clear problem with taking this approach is the summation over $|V|$ exponents present in the loss function. For large collections of text, $|V|$ may be several hundreds of thousands. Therefore, a computational optimisation technique known as negative sampling is often used, in which we avoid the summation over $|V|$ but instead sum over K negative samples, taken randomly from ‘noisy’ negative samples selected randomly from the vocabulary (Mikolov et al.,

2013b). The values taken for K typically range from 2 to around 20, with smaller training sets allowing for larger values of K . In our application, we shall be considering an extension of the skip-gram negative sampling (SGNS) model known as FastText.

3.2 FastText

Word2Vec is widely considered to be a state-of-the-art word embedding language model. However, one of its main flaws is that it does not make use of sub-words. For example, the words “walk” and “walking” would be considered as distinct words, and the fact that the two words have the same stem, namely “walk”, is not used when learning the two vector representations. This problem becomes apparent also with misspelled words, as Word2Vec will treat such words as distinct from their correctly-spelled counterparts.

FastText (Bojanowski et al., 2017) is an approach based on the Word2Vec SGNS model that also considers sub-word information. It does so by learning representations for character n -grams rather than for individual words. As the name implies, a character n -gram is a sequence of n characters obtained from a particular word. For example, given the word “gravity”, the set of character 3-grams is given by

$$\{ \langle \text{gr, gra, rav, avi, vit, ity, ty} \rangle \},$$

where \langle and \rangle are start- and end-of-word markers respectively.

In practice, each word w_i is represented as a bag of character n -grams $\mathcal{G}_{w_i, n_{min}, n_{max}}$, where n is allowed to vary from a pre-specified minimum value n_{min} to a pre-specified maximum value n_{max} . The word itself is also included in this set as a special sequence. Different values for n_{min} and n_{max} can be considered. However, the values we shall be using in our application are the same as in the original study by Bojanowski et al. (2017); i.e., $n_{min} = 3$ and $n_{max} = 6$.

A particular strength of FastText is its ability to deal with compound words; that is, words that are composed of multiple smaller words. Additionally, FastText can identify prefixes and suffixes, and can infer the meaning of “unseen” (i.e. out-of-vocabulary) words based on the surrounding context (Bojanowski et al., 2017). Similarly, FastText recognises a misspelled word as being close to its correctly-spelled counterpart due to the significant overlap in character n -grams.

3.3 Doc2Vec

While FastText is a state-of-the-art algorithm that may be used to learn word embeddings, the need may arise

to obtain embeddings corresponding to documents, called document embeddings, rather than individual words. Within the context of our application, each newspaper comment will be considered as a separate document, and a vector embedding corresponding to each document will be obtained. Such embeddings can be obtained by means of extensions to the Word2Vec framework which cater for whole documents.

While more sophisticated extensions, such as Paragraph Vector (Le and Mikolov, 2014), exist, we shall consider a simpler approach. In particular, we consider an R programming language implementation of a Word2Vec extension known as Doc2Vec, which may be found in the R package `textTinyR` (Mouselimis, 2019). While the Paragraph Vector models construct document embeddings from scratch given the one-hot encoded vectors as input in a manner similar to the two Word2Vec models, the Doc2Vec algorithm takes pre-learned word embeddings and combines them to form document embeddings. Three different methods can be implemented in R in order to convert word vectors to document vectors, namely `sum_sqrt`, `min_max_norm` and `idf`. Of these, we shall be using the `sum_sqrt` method. `idf` refers to term frequency - inverse document frequency (commonly known as TF-IDF). On our dataset, TF-IDF did not produce meaningful results, possibly due to the short and unstructured nature of our documents (i.e., news comments). On more structured and lengthier documents, like e.g. *résumés* (Grech and Suda, 2020), this has the potential of giving better results.

4 APPLICATION AND RESULTS

This section deals with the application of the mentioned techniques to a dataset of online newspaper comments on articles found in the ToM website. The dataset was obtained from the Department of Linguistics at the University of Malta, who obtained it directly from ToM. All comments in the dataset were written between April 2008 and January 2017. Currently, the Department of Linguistics is carrying out a study on the same dataset regarding hate speech and critical discourse analysis (Assimakopoulos et al., 2020).

The dataset contains a total of 2,141,090 comments. The analysis was conducted only on English-language comments, of which there were 1,815,965. These comments vary greatly in length; some contain a single word or a short phrase, while others contain multiple lengthy paragraphs. The average length of these comments is 22.25 words. Extensive clean-

ing and pre-processing of the data was carried out. The excluded comments were mainly in Maltese or contained significant code-switching; that is, alternating between multiple languages. Code-switching is a very common phenomenon in Maltese online comments (and also verbal conversation) due to the fact that Malta is largely fluently bilingual (Maltese and English), if not trilingual in some cases (usually Italian being the third language). Detection of non-English comments was done using the `clD2` package in R. Comments which were deleted by the ToM's moderators are also excluded from the dataset. As per the ToM's comment policy (Times of Malta, 2020), comments may be deleted if they are considered to be defamatory, racist, sexist, or otherwise offensive.

We started the analysis by looking at the most common n -grams (up to $n = 4$), excluding stop words. These indicate that the overall content of the dataset is highly political in nature. Some overtly political references found within the most frequent n -grams include references to “government”, “prime minister”, and names of politicians such as the names of three former Maltese Prime Ministers, namely Dr Joseph Muscat, Dr Lawrence Gonzi, and Dr Eddie Fenech Adami. References are also made to politicised issues such as the national airline Air Malta and the new power station. The absence of vocabulary related to arts, sports and culture is conspicuous. Argumentative language is also commonly used and seems to crop up most commonly in 4-grams; this includes expressions such as “two weights two measures” and “pot calling kettle black”. Passive-aggressive phrases such as “time will tell” and backhanded compliments such as “well done” and “keep good work” also appear frequently in the data.

4.1 Embeddings and Clustering

Word vectors were extracted for each word in the vocabulary, that is, for each word present in the comments. The FastText word embedding model was used for this task. Stop words were removed for this analysis as they are not expected to contribute anything meaningful in determining the topic of each comment. While stemming (i.e., removing the suffix of a word to reduce it to its root) was not carried out due to computational difficulties, the effect of stemming is still largely achieved. For example, the embeddings for the words “worked” and “working” should be similar since there is a significant overlap in the bags of character n -grams for the two words. For clustering arising from words or documents, we only present the results of the most recent period of data available, i.e. January 2016 until January 2017.

Table 1: Sample of 5 words from each cluster: $k = 20$, 2016 and January 2017 data, and corresponding identified topics.

Cluster	Sample Words	Identified Topics
1	caught, done, denied, sold, expected	Past tense
2	high, yet, full, otherwise, proposal	No clear topic
3	applied, owned, planned, agreed, died	Past tense
4	pregnancy, cannabis, divorce, priests, gay	Religion; Civil Liberties
5	question, education, nation, taxation, application	Mostly words ending in -ion or -ions
6	shipyard, billboards, wayward, cardboard, laggards	Mostly words ending in -ard or -ards
7	laburist, barranin, partit, gvern, ajkla	Maltese words related to politics
8	gonzi, scicluna, eddy, alfred, marlene	Maltese politicians' names and surnames
9	crime, freedom, sentence, protest, partisan	Politics; Justice
10	offshore, papers, scandal, resign, taxes, corruption	Corruption; Scandals
11	uk, eu, migrants, brexit, deport	Foreign Politics/Brexit; Immigration
12	power, water, bills, solar, interconnector	Energy
13	just, like, little, news, happens	No clear topic
14	trump, christians, islamic, refugees, extremist	Religion; Immigration
15	government, police, citizens, party, hypocrites	Local Politics
16	vehicle, infrastructure, licence, wardens, cyclists	Transport; Traffic
17	going, meeting, breaking, growing, eating	Words ending in -ing
18	hundred, percentage, billions, fee, debt	Finance; Numbers
19	properties, ugly, paceville, hotels, visitors	Tourism; Construction
20	compete, comparing, comedy, commercial, column	Mostly words starting with co- or com-

The hyperparameter values used for FastText are as follows: $\text{learning_rate} = 0.025$, $\text{learn_update} = 100$, $\text{word_vec_size} = 300$, $\text{window_size} = 5$, $\text{epoch} = 5$, $\text{min_count} = 3$, $\text{neg} = 5$, $\text{min_ngram} = 3$, $\text{max_ngram} = 6$, $\text{nthreads} = 8$, $\text{threshold} = 0.0001$.

After the word embeddings were obtained, k -means clustering on these vectors was carried out. However, the usual methods for determining the optimal value of k , such as the elbow method, proved inconclusive. Multiple values of k were tried. Of these, $k = 20$ arguably provided the most interpretable results. Lowering the value of k , for example $k = 10$, results in some important topics not being assigned a cluster. On the other hand, increasing k , say to $k = 30$, results in some topics being spread across multiple clusters. As an example, a sample of 5 words from each cluster for $k = 20$ are shown in Table 1, along with the identified topic. The same exercise is repeated for the document embeddings, again taking $k = 20$ for the same reasons outlined above. A sample comment from each of the $k = 20$ clusters is shown in Table 2, along with the identified topic. In this case, certain comments need to be understood within a local context, such as the 'Cafe Premier' mention in the Cluster 2 comments, which makes reference to a specific political scandal, and the "make hay while the sun shines" comment made by a Maltese construction magnate in Cluster 5 which went viral. It is also in-

teresting to note that positive and sarcastic comments tend to be clustered together (Cluster 14).

To examine the effect of clustering when considering different time periods, extraction and clustering of word and document vectors was also carried out on different subsets of the dataset. The different subsets of data considered are as follows 2008-2011, 2012, 2013, 2014, 2015, 2016 and January 2017. While most topics were observed to be prevalent throughout all time periods (such as civil liberties, construction, religion and economy/finance), a number of topics stood out in certain years. For example, divorce was a highly-debated topic for the comments written in 2011 or earlier. This is due to the widespread national debate on the topic which ultimately led to a referendum in Malta in 2011, following which divorce was legalised. In addition, the introduction of the Individual Investor Programme citizenship scheme in 2013 led to much debate on the nature of Maltese citizenship throughout that year. It is evident that topics that can polarize opinions generate more interest and discussion from readers. A topic which stood out in 2014 is Russia's annexation of Crimea and the resulting conflict with Ukraine. This shows that it is not only local topics that are discussed within online newspaper comments on the Times of Malta, but also international ones. Similarly, the presence of a cluster of word vectors for 2015 corresponding to words related

Table 2: Sampled comment from each cluster: $k = 20$, 2016 and January 2017 data. Note that the comments have had stopwords removed, and are hence not presented in a grammatically correct form.

Cluster	Sample Comment	Identified Topics
1	curves round bouts one can hardly achieve speed 80kph	Transport
2	nothing complain living cuckoo land scandals surfacing since day one remember cafe premier	Corruption
3	thing learn africa lost another 366 persons produced something countries improve lives rest population	Foreign Politics
4	thought know now promises minister just hot air unless concerns family right wrong	No clear topic
5	ones supplied concrete now making hay whilst sun shines want rock boat	Construction
6	revolution tenderness pope francis spoke may god bless protect pope francis	Religion
7	voters spoken allow freedom trump enacting commands electorate actually simple americans fed	Foreign Politics
8	says majority behind pn source even change relation news item	Local Politics
9	appointment liberal wing np de marco gaining ground confessional wing fenech adami	Local Politics
10	mr t***** unwanted pregnancy can never solved killing innocent defenceless child fundamental human right life also irresponsible sex carry right commit murder	Abortion; Alcohol & Drugs; Civil Liberties
11	good points last 25 years informed public registry system dated back 1992 serviced single person passed away	Local Politics
12	pity activists can protest around rest islands hotels taken privatised lot areas immediately front hotels shores argument	Tourism; Construction
13	really think legal notice safeguard road safety dejquhom [sic] il billboards ta panamagate	Comments containing code-switching
14	wow news value father karl marx monarchist mother bernard shaw conservative margaret tatcher father	Positive comments; Sarcastic comments
15	mizzi schembri financial affairs strictly confidential non existant [sic] vat receipts also confidential	Panama Papers; Corruption
16	norway finland legal world laws within european union legal personal use includes malta	European Politics
17	can say likes work malta also will tram system due cost maintenance local cultural issues	Local Economy
18	really excuse accidents nature however problem will remain since government police courts fail address	Law & Justice
19	politicians disgusting totally idiotic	Agreement/Disagreement
20	dockyard many many years fuss now past providing workers many workers contributing malta economy	Local Economy

to the United Kingdom will be due to Malta's hosting of the Commonwealth Heads of Government Meeting during that year, as well as due to the ever-increasing discussion around Brexit. Finally, two hotly debated topics during the period January 2016-January 2017 were the morning-after pill, a topic which proved controversial with people holding pro-life views, as well

as the Panama Papers and related corruption allegations involving high-ranking politicians. These allegations significantly altered the political landscape of the country, and arguably led to the announcement of a snap election in June 2017. Sports-related clusters, on the other hand, appear in 2012-2014, likely due to the Euro Cup, the Olympics and the World Cup.

Table 3: Topics identified through k -means clustering for each time period.

Time Period	Identified Topics
2016 and January 2017	Abortion; Civil Liberties; Construction; Corruption; Drugs; Economy & Finance; Education; Energy; Environment; European Politics; Foreign Politics; Hunting; Immigration; Law & Justice; Local Politics; Morning-After Pill; Panama Papers; Religion; Tourism; Transport & Traffic
2015	Abortion; Brexit & UK Topics; Civil Liberties; Construction; Economy & Finance; Education; Environment; European Politics; Foreign Politics; Hunting; Immigration; Law & Justice; Local Councils; Local Politics; Religion; Science; Transport & Traffic
2014	Alcohol & Tobacco; Civil Liberties; Construction; Drugs; Economy & Finance; Education; Employment; Energy; Environment; Foreign Politics; Human Rights; Hunting; Immigration; Law & Justice; Local Politics; Medicine; Religion; Russia, Ukraine & Crimea; Science; Sports & Culture; Transport & Traffic
2013	Abortion; Civil Liberties; Construction; Economy & Finance; Energy; Environment; European Politics; Foreign Politics; Hunting; Immigration; Law & Justice; Local Politics; Maltese Citizenship; Religion; Sports & Culture; Tourism; Transport & Traffic
2012	Abortion; Civil Liberties; Construction; Divorce; Economy & Finance; Education; Environment; European Politics; Feasts and Festivals; Foreign Politics; History; Hunting; Immigration; Law & Justice; Local Politics; Medicine; Religion; Science; Sports & Culture; Technology & Media; Tourism; Transport & Traffic; Travel
2008-2011	Abortion; Alcohol & Tobacco; Animal Welfare; Armed Forces; Civil Liberties; Construction; Divorce; Drugs; Economy & Finance; Education; Energy; Environment; European Politics; Foreign Politics; Hunting; Immigration; Law & Justice; Local Politics; Religion; Tourism; Transport & Traffic; Travel

A summary of the different topics identified for each subset of the data considered is given in Table 3.

5 CONCLUSION

Neural word embedding models such as Word2Vec, FastText, and Doc2Vec are effective tools for providing vector representations of words and documents. In particular, these tools have been applied to a dataset of online newspaper comments, where each comment was taken to be a separate document.

The first part of the application was concerned with cleaning and pre-processing the data, and then obtaining descriptive statistics in order to understand the data better. The second part of the application section considered the implementation of word embedding models to the online newspaper comments dataset. FastText was used to generate word embeddings, which were then grouped into clusters. These word embeddings were then fed into Doc2Vec in order to produce document embeddings, and the clustering exercise was repeated on the document vectors. For each time period considered, a number of topics

were identified as being more prevalent than others during that time period.

The main limitation encountered was the computational intensiveness of the data analysis. In addition, more recent data (later than January 2017) would have presented a more contemporary picture of what piques interest from a Maltese online news portal audience. It should also be noted that the use of k -means clustering might have presented problems, especially since we dealt with high-dimensional data, and perhaps other clustering algorithms and alternative distance metrics (Aggarwal et al., 2001) or more sophisticated methods such as Latent Dirichlet Allocation (Jacobi et al., 2016) could have been used for topic modelling instead.

Related and further possible research work in this area may include solving word analogies (Mikolov et al., 2013b), bias detection (Bolukbasi et al., 2016), applying the Joint Topic-Expression model (Mukherjee and Liu, 2012; Liu, 2015), and the use of recursive neural networks for tasks such as sentiment analysis (Socher et al., 2011; Socher et al., 2013b), sentence parsing (Socher et al., 2013a), and political ideology detection (Iyyer et al., 2014).

ACKNOWLEDGEMENTS

The authors would like to thank Dr Albert Gatt for allowing the use of a GPU server. We would also like to thank Dr Lonneke van der Plas, Dr Stavros Assimakopoulos and Ms Rebekah Vella Muskat for providing the dataset used in this study.

REFERENCES

- Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *ICDT '01*, pages 420–434, London, United Kingdom.
- Allison, B., Guthrie, D., and Guthrie, L. (2006). Another look at the data sparsity problem. In *TSD '06*, pages 327–334, Brno, Czech Republic.
- Assimakopoulos, S., Vella Muskat, R., van der Plas, L., and Gatt, A. (2020). Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *LREC '20*, pages 5088–5097, Marseille, France. ELRA.
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *J Mach Learn Res*, 3(Feb):1137–1155.
- Bengio, Y. and LeCun, Y. (2007). Scaling learning algorithms towards AI. In Bottou, L., Chapelle, O., DeCoste, D., and Weston, J., editors, *Large scale kernel machines*. MIT Press, Cambridge, MA.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans Assoc Comput Linguist*, 5:135–146.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS '16*, pages 4349–4357, Barcelona, Spain.
- Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n -gram models of natural language. *Comput Linguist*, 18(4):467–480.
- Costola, M., Nofer, M., Hinz, O., and Pelizzon, L. (2020). Machine learning sentiment analysis, COVID-19 news and stock market reactions. *SAFE Working Paper*.
- Grech, B. and Suda, D. (2020). A neural information retrieval approach for résumé searching in a recruitment agency. In *ICPRAM '20*, pages 645–651, Valletta, Malta. SciTePress Digital Library.
- Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). A closer look at skip-gram modelling. In *LREC '06*, Genoa, Italy. ELRA.
- Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *ACL-IJCNLP '14, Volume 1*, pages 1113–1122, Baltimore, MD.
- Jacobi, C., Van Atteveldt, W., and Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digit Journal*, 4(1):89–106.
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML '14*, pages 1188–1196, Beijing, China.
- Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, chapter 10: Analysis of Debates and Comments, page 231–249. Cambridge University Press, Cambridge, UK.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *ICLR '13*, Scottsdale, AZ.
- Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *INTERSPEECH '10*, Makuhari, Japan.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS '13*, pages 3111–3119, Stateline, NV.
- Mouselimis, L. (2019). textTinyR: Text processing for small or big data files. Retrieved from <https://CRAN.R-project.org/package=textTinyR> on September 22, 2020.
- Mukherjee, A. and Liu, B. (2012). Mining contentions from discussions and debates. *ACM-SIGKDD '12*, pages 841–849.
- Pibiri, G. E. and Venturini, R. (2019). Handling massive n -gram datasets efficiently. *ACM Trans Inf Syst*, 37(2):1–41.
- Pickhardt, R., Gottron, T., Körner, M., Wagner, P. G., Speicher, T., and Staab, S. (2014). A generalized language model as the combination of skipped n -grams and modified Kneser-Ney smoothing. *arXiv:1404.3377*.
- Rădulescu, C., Dinsoreanu, M., and Potolea, R. (2014). Identification of spam comments using natural language processing techniques. In *ICCP '2014*, pages 29–35. IEEE.
- Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013a). Parsing with compositional vector grammars. In *ACL-IJCNLP '13, Volume 1*, pages 455–465, Nagoya, Japan.
- Socher, R., Lin, C. C., Manning, C. D., and Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *ICML '11*, pages 129–136, Bellevue, WA.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., et al. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP '13*, pages 1631–1642, Seattle, WA.
- Times of Malta (2020). Comment policy - Times of Malta. Retrieved from <https://timesofmalta.com/comments> on September 24, 2020.
- Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Comput Linguist*, 40(1):171–202.