Neural Topic Modeling in Social Media by Clustering Latent Hashtag Representations

Riccardo Cantini, Cristian Cosentino, Fabrizio Marozzo, Domenico Talia, Paolo Trunfio

DIMES, University of Calabria, Rende, Italy

28th European Conference on Artificial Intelligence (ECAI–2025)

October 29, 2025 || Bologna, Italy

Motivations

- **Social media** has become a major space for public discourse, providing valuable data to study social dynamics, emerging trends, and public sentiment.
- Extracting **meaningful topics** from social media data remains difficult due to its informal, dynamic, and context-dependent nature.
  - Posts are short, noisy, and highly tied to real-time events, which complicates the identification of coherent topic structures.
  - Traditional topic models, such as LDA and pLSA, rely on word co-occurrence patterns in structured text, failing to handle the **brevity** and **noise** of social media.
  - Neural Topic Models (NTMs) based on pre-trained encoders, such as BERTopic and Top2Vec, often fail to capture **domain-specific** and **rapidly evolving** meanings.

Main challenge: effectively model dynamic, context-sensitive language in social media, capturing **highly localized nuances** of meaning.

The Role of Hashtags

- **Hashtags** are often overlooked in traditional topic modeling, treated like any other word in social media corpora.
- However, they have a unique semantic role:
  - They convey **topical information**, grouping posts with similar content while tying them to trending topics, trends, and sentiment.
  - They reflect **emerging trends** and **evolving discourse**, capturing real-time shifts in social media conversations.
  - They provide a **compact signal** for topic modeling, summarizing the essence of short and noisy posts.

Main idea: leverage **hashtags as semantic anchors** to learn corpus-specific topic structures, addressing the evolving and context-sensitive nature of social posts.

Proposed Methodology: NTM-HEC

NTM-HEC (*Neural Topic Modeling via Hashtag Embedding Clustering*) is a **hashtag-centric** framework for **neural topic modeling**.

- It builds on the **modular design** of NTMs like BERTopic and Top2Vec, but adapts to domain-specific discourse by focusing on hashtags.
- It treats hashtags as meaningful **proxies for the topic-related semantics** of social media posts.
- It learns **corpus-specific hashtag embeddings** directly from the target dataset of social posts, ensuring sensitivity to highly localized and contextual nuances.

NTM-HEC grounds topic discovery in learned hashtag semantics, overcoming NTMs that rely on generic pre-trained encoders and may miss local or evolving meanings.

Illustrative Example: Domain-Specific Semantics with NTM-HEC

- Consider the hashtags *#Azov* and *#AzovBattalion*, frequently used on *X* during the early months of the **Russia-Ukraine conflict**.
- Standard NTMs using off-the-shelf embeddings (e.g., Sentence-BERT) might cluster these posts under **general military topics** due to surface-level similarities.
- By learning from co-occurrence patterns in the specific discourse, NTM-HEC identifies these hashtags as central to a finer-grained topic about:
  - the **siege of Mariupol**
  - the **surrender of the Azov Battalion**, who were barricaded inside the Azovstal steel plant in Mariupol.
- NTM-HEC clusters these hashtags with *#saveAzov*, *#saveMariupol*, and *#zelensky*, capturing **temporal specificity** and **ideological framing** that would likely be missed by generic pre-trained encoders.

NTM-HEC Pipeline Overview

Main steps:

1. **Learning corpus-specific hashtag embeddings**: train a CBoW Word2Vec model on the target corpus to embed words and hashtags jointly.
2. **Low-dimensional projection of hashtag embeddings**: apply t-SNE (PCA-initialized) to obtain a 2D representation of latent hashtag representations.
3. **Hashtag clustering for topic discovery**: use HDBSCAN to group semantically similar hashtag projections into coherent and interpretable topics.

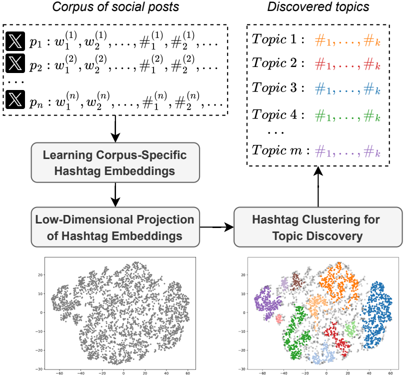


Figure: Execution flow of NTM-HEC.

Step 1 – Learning Corpus-Specific Hashtag Embeddings

- Train a **Continuous Bag-of-Words (CBoW)** Word2Vec model directly on the target social media corpus.
- Jontly embed words and hashtags in a **shared semantic space**, capturing their co-occurrence context.
  - Hashtag meaning emerges from surrounding words, producing **semantically grounded representations**.
  - Example: *#NoFlyZone* → embedding shaped by "*airspace*", "*NATO*", "*conflict*", aligning with military and geopolitical context.
- The final Word2Vec model provides domain-specific embeddings that reflect **localized**, **context-specific** hashtag semantics.

Step 2 – Low-Dimensional Projection of Hashtag Embeddings

- All words and hashtags are embedded into a 150-dimensional latent space using the **Word2Vec model**.
- Retain only **hashtag embeddings**, as they capture topical information.
- Apply **dimensionality reduction** using PCA-initialized **t-SNE** to project embeddings into 2D space:
  - Improves **stability** of hashtag-based clustering structures.
  - Enhances **interpretability** of topic clusters and enables **visualization** of hashtag topology and semantic relationships.
  - **Barnes-Hut approximation** preserves local neighborhood structure and global clustering patterns while ensuring **scalability**.

Step 3 – Hashtag Clustering for Topic Discovery

- Cluster low-dimensional hashtag embeddings to identify **coherent topic groups**.
- Why **HDBSCAN**?
  - Detects clusters of **varying shapes** due to its density-based approach.
  - Adapts to **variable topic densities**, capturing both macro-topics and micro-trends without requiring a fixed number of topics.
  - Naturally **filters out outliers** and noise hashtags that do not belong to coherent topical groups.
- Topic clusters:
  - HDBSCAN produces **coherent**, **interpretable**, and **non-overlapping** topic clusters.
  - Each cluster represents a distinct **discussion topic**, summarized by its top-k (most frequent) hashtags.

Experimental Evaluation

- **Case Studies**:
  - **Russia–Ukraine Conflict** (Mar–Jun 2022)
  - **COVID-19 Pandemic** (Dec 2020–Mar 2021)
- **Datasets**:
  - Publicly available corpora of *X* posts written in English
  - **Russia–Ukraine:** 100K tweets/month
  - **COVID-19:** 400K tweets/month
  - Enables analysis of both **long-term dynamics** and **short-term trends**
- **Evaluation Metrics**:
  - **Topic Coherence:** *CV*, *NPMI* — semantic consistency of top hashtags
  - **Topic Diversity:** *PUW*, *Jaccard Distance* — non-redundancy of discovered topics
  - **Embedding-based:** SIL_{PW}, SIL_{CB} — cohesion and separation in embedding space

Exploratory Analysis of NTM-HEC Configuration

- We compared different configurations for each step of the NTM-HEC pipeline to identify the most effective combination.
  - **Hashtag Embedding**: Word2Vec vs. FastText
  - **Dimensionality Reduction**: t-SNE, UMAP, or none
  - **Clustering**: HDBSCAN, K-Means, Gaussian Mixture Models (GMM)
- **Results Overview**:
  - **Word2Vec** outperforms FastText, showing better semantic consistency.
  - **t-SNE** yields more stable and coherent clusters than UMAP or no reduction.
  - **HDBSCAN** surpasses K-Means and GMM in all considered metrics.

Step	Alternative	CV	NPMI	PUW	JD	SIL_CB	SIL_PW
Hashtag Embedding	FastText	0.473 ± 0.039	0.013 ± 0.011	1.000 ± 0.000	1.000 ± 0.000	0.403 ± 0.069	0.591 ± 0.051
Dimensionality Reduction	None	0.411 ± 0.036	-0.031 ± 0.047	1.000 ± 0.000	1.000 ± 0.000	0.437 ± 0.113	0.628 ± 0.069
	UMAP	0.462 ± 0.044	0.029 ± 0.012	1.000 ± 0.000	1.000 ± 0.000	0.529 ± 0.072	0.665 ± 0.042
Hashtag Embedding Clustering	K-Means	0.433 ± 0.055	0.014 ± 0.010	1.000 ± 0.000	1.000 ± 0.000	0.503 ± 0.078	0.637 ± 0.052
	GMM	0.436 ± 0.050	0.028 ± 0.017	1.000 ± 0.000	1.000 ± 0.000	0.510 ± 0.070	0.641 ± 0.031
Reference configuration W2V + t-SNE + HDBSCAN		**0.500** ± 0.048	**0.068** ± 0.036	**1.000** ± 0.000	**1.000** ± 0.000	**0.560** ± 0.061	**0.721** ± 0.047

Discovered Topics – Russia–Ukraine Conflict

- Analysis of 400k tweets (March–June 2022) focused on online discussions surrounding the Russia–Ukraine war.

- **Long-term Topics**:
  - **War zones & cities**: #Kyiv, #Kharkiv, #Donbass, #Mariupol, #Odessa, ...
  - **Pro-Ukraine discourse**: #StandWithUkraine, #SlavaUkraini, #StopRussia, ...
  - **Pro-Russia narratives**: #IStandWithRussia, #NaziUkraine, #AbolishNATO, ...

- **Short-term / Event-driven Topics:**
  - **Operation Ganga**: evacuation of Indian citizens (Mar 2022).
  - **Battle of Donbas**: escalation in eastern Ukraine (Apr 2022).
  - **Eurovision 2022**: cultural solidarity, victory of Kalush (May 2022).
  - **Azov Battalion**: siege and surrender at Mariupol's Azovstal plant (May 2022).
  - **Economic Forum**: WEF annual meeting in Davos, Switzerland (May 2022).
  - **G7 Summit 2022**: geopolitical focus on China and sanctions (Jun 2022).

NTM-HEC captures both **persistent narratives** (war, geopolitics) and **emerging micro-events**, demonstrating sensitivity to temporal and ideological nuances in online discourse.

Discovered Topics – COVID-19 Pandemic

- Analysis of 1.6M tweets (Dec 2020 – Mar 2021) about vaccination and pandemic management.

- **Long-term Topics:**
  - **Pharmaceutical industry**: #AstraZeneca, #Pfizer, #Johnson, ...
  - **US politics**: #Biden, #Trump, #OperationWarpSpeed, ...
  - **Public health measures**: #Lockdown, #Quarantine, #SocialDistancing, ...
  - **Vaccine safety**: #VaccineAllergy, #Anaphylaxis, #VaccineSideEffects, ...
  - **Conspiracy narratives**: #MicrochipVaccines, #5GConspiracy, #BillGates, ...

- **Short-term / Event-driven Topics**:
  - **Cyberattack on EMA**: breach of the European Medicines Agency (Dec 2020).
  - **China vaccine donation to Pakistan**: 500k Sinopharm doses (Jan 2021).
  - **Second wave**: rising fears over a second wave of Covid infections (Jan 2021).
  - **Vaccine distribution**: focus on equitable global access to vaccines (Feb 2021).
  - **School reopening**: debate over the pandemic's impact on education. (Feb 2021).
  - **AstraZeneca concerns**: reports of rare blood-clotting events linked to the vaccine, with temporary restrictions in several countries (Mar 2021).

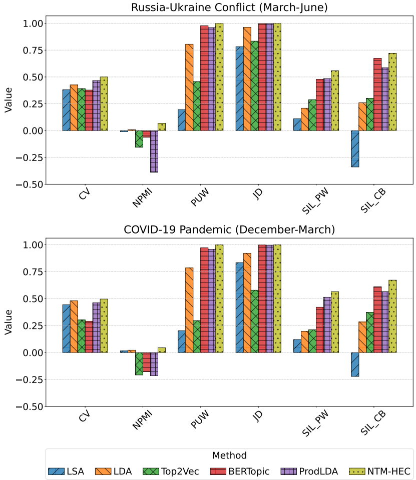
State-of-the-art Comparison

Competing Methods:

- Traditional: **LSA**, **LDA**
- Neural: **Top2Vec**, **BERTopic**, **ProdLDA**

Results Summary:

- **Higher topic coherence** (CV, NPMI), indicating superior topic quality and interpretability.
- **Maximum diversity** (PUW, JD) due to non-overlapping hashtag clusters.
- **Superior embedding-based scores** (SIL_{CB}, SIL_{PW}), indicating semantically cohesive, well-separated topics.
- Consistent improvement over classical and neural baselines, by harnessing **corpus-specific hashtag embeddings**.



Conclusions and Future Work

- We introduced **NTM-HEC**, a hashtag-centric neural topic modeling framework for context-aware topic discovery.
  - Combines **hashtag embedding**, **manifold learning**, and **clustering** to uncover coherent, diverse, and interpretable topics.
  - Captures both **persistent themes** and **emerging events** in social media discourse, effectively grasping **context-specific**, **highly-localized** nuances of meaning.
- **Main Findings**:
  - Outperforms **traditional** (LDA, NMF) and **neural** (BERTopic, Top2Vec, ProdLDA) approaches across all datasets.
  - Achieves higher topic **coherence**, **diversity**, and **semantic separability**.
- **Future Directions**:
  - Extend NTM-HEC to **multilingual** and **multimodal** settings.
  - Integrate **temporal modeling** to track topic evolution over time.
  - Explore **hybrid topic models** combining transformer encoders with corpus-specific hashtag representations.

Any questions?

rcantini@dimes.unical.it